

**„DIE KLASSENAUSWAHL IN
MUSTERERKENNUNGSPROZESSEN
DARGESTELLT AM BEISPIEL COMPUTER-GESTÜTZTER
DIAGNOSEVERFAHREN IN DER MEDIZIN“**

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaft

am Fachbereich Mathematik
der Johannes Gutenberg-Universität in Mainz

Friedrich Heß
geb. in Koblenz

Mainz 1979

**„DIE KLASSENAUSWAHL IN
MUSTERERKENNUNGSPROZESSEN
DARGESTELLT AM BEISPIEL COMPUTER-GESTÜTZTER
DIAGNOSEVERFAHREN IN DER MEDIZIN“**

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaft

am Fachbereich Mathematik
der Johannes Gutenberg-Universität in Mainz

Friedrich Heß
geb. in Koblenz

Mainz 1979



Berichterstatter:

Prof. W.Bühler, Ph.D., Mainz

Prof. Dr. U.Kuss, Mainz

Datum der mündlichen Prüfung: 20.12.1979

INHALT

EINLEITUNG

KAPITEL I

DIE BAYES ENTSCHEIDUNG	1
§0 Generelle Voraussetzungen und Bezeichnungen	1
§1 Änderung der Klasseneinteilung	2
§2 Abschätzungen des Bayes - Risikos	8
§3 Informationsgehalt einer Klasseneinteilung	17
§4 Gütekriterien für die Klassenauswahl	23
§5 Einflüsse der Verlustfunktion	27
§6 Einflüsse der a priori Verteilung	31
§7 Numerische Ergebnisse	39

KAPITEL II

DIE REJECT ENTSCHEIDUNG	47
§1 Reject- und Fehlklassifizierungsrisiko	47
§2 Spezielle Verlustfunktionen	50
§3 Error-Reject-Tradeoff	53

KAPITEL III

DIE NEAREST NEIGHBOR ENTSCHEIDUNG	58
-----------------------------------	----

KAPITEL IV

MODIFIZIERTE ENTSCHEIDUNGSREGELN	63
§1 Modifizierte Bayes - Entscheidung	63
§2 Modifizierte Reject - Entscheidung	69
§3 Modifizierte Entscheidung mit sequentieller Merkmalsauswahl	78
Liste der Bezeichnungen	82
LITERATURVERZEICHNIS	84



Digitized by the Internet Archive
in 2026

EINLEITUNG

In der Theorie der Mustererkennung werden i.a. folgende Schritte behandelt (vgl. [3], [39], [45]):

- (1) Wahl problemorientierter Merkmale und Bestimmung der zu messenden Größen ("feature selection")
- (2) Wahl der Klassen, zwischen denen eine Unterscheidung vorgenommen werden soll
- (3) Wahl einer geeigneten Klassifizierungsregel mit Schätzung der zugrunde liegenden Verteilungen im parametrischen Fall
- (4) Durchführung der Zuordnung

Da die Schritte (1) - (3) in zahlreichen Wechselwirkungen zueinander stehen ([3], [12], [39], [44], [45]), werden sie in praxi in verschiedenen Reihenfolgen angewandt. Dabei gibt es besonders für die Auswahl von Merkmalen und für die Wahl der Klassifikationsregel - hier insbesondere für die Schätzung der zugehörigen Verteilungen - eine Vielzahl von Theorien und praktischen Verfahrensregeln ([3], [12], [28], [29], [30], [31], [33], [39], [46]).

Dagegen wird die Auswahl der zu unterscheidenden Klassen meist nur anhand pragmatischer und heuristischer Kriterien ausgeführt. In automatischen - d.h. computergestützten - Diagnoseverfahren in der Medizin beinhaltet diese Frage so - wohl den Umfang der zu behandelnden Differentialdiagnosen als auch das Problem der Bildung von Subklassen und damit der weiteren Differenzierung bekannter Krankheiten. Dies ist u.a. von PIPBERGER et al. [36] für die Differentialdiagnose des "Brustschmerzes" diskutiert worden (vgl. auch [31], [19]).

Das Ziel meiner Arbeit ist nun, quantitative Kriterien für die Klassenauswahl zu schaffen, die einerseits den pragmatischen Forderungen weitgehend gerecht werden und andererseits im Rahmen der mathematischen Entscheidungstheorie bestimmt und diskutiert werden können.

Eines dieser Kriterien muß zwangsläufig die entscheidende Größe jedes Mustererkennungsprozesses, das Fehlklassifizierungsrisiko, beinhalten. Während dieses mit sinkender Klassenanzahl i.a. abnimmt, verringert sich gleichzeitig die Unterscheidungskraft oder der Informationsgehalt der Klasseneinteilung. Es ist also notwendig, auch für die Information, die eine gegebene Klassenauswahl bereithält, eine meßbare Größe zu definieren. Dazu bemerkten GOOD und CARD [14]:

"No measure of loss due to error of observation can be found without first selecting a measure of the value of information obtained without errors of observation". Dabei gilt es dann, einen (optimalen) Kompromiß zwischen beiden Forderungen, einer möglichst feinen Klasseneinteilung und einem großen Informationsgehalt einerseits und einer geringen Irrtumsrate andererseits zu finden.

Da die Irrtumsrate offenbar von der Auswahl der Merkmale und der Wahl der Klassifizierungsregel abhängt, kann das Problem einer optimalen Klassenauswahl nur bei fester Wahl dieser beiden anderen Punkte diskutiert werden. Deshalb ist diese Arbeit nach der Wahl der Entscheidungsverfahren in vier Kapitel unterteilt. In den ersten drei Kapiteln werden bei fester Merkmalsauswahl Gütekriterien für die Klasseneinteilung für zwei parametrische Verfahren - die Bayes-Entscheidung

(Kap.I) und die Reject-Entscheidung (Kap.II) - und für die nicht parametrische "Nearest-Neighbor"-Entscheidung behandelt. In Kap.IV werden dann neue, modifizierte Entscheidungsverfahren entwickelt, die auch eine variable Handhabung der Merkmalsauswahl ermöglichen.

Bei der Festlegung auf die genannten Klassifizierungsregeln diente mir als Basis die Arbeit, in der DAS GUPTA [11] die verschiedenen Klassifikationsverfahren vergleichend behandelt hat. Um die für medizinische Diagnoseverfahren relevanten Regeln herauszufinden, untersuchte ich die Arbeiten der von WAGNER et al.[43] herausgegebenen Bibliographie über medizinische Diagnoseprobleme. Daneben diente mir die Habilitationsschrift von MICHAELIS [31] als Richtschnur. Wegen der Bedeutung der Ekg-Analyse innerhalb der automatischen Diagnoseprozeduren nimmt diese einen besonderen Platz bei der Anwendung der hier erzielten Ergebnisse ein. So werden die Resultate fast ausschließlich am Beispiel der multivariaten Normalverteilung erläutert.

In KAPITEL I wird die Bayes-Entscheidung (BE) behandelt. Sie nimmt in dieser Arbeit den größten Raum ein, da sie einerseits in der Entscheidungstheorie am besten untersucht ist und da sie sich andererseits in ihrer Anwendung in Diagnoseverfahren als sehr geeignet erwiesen hat. So brachte ihre Anwendung in der von PIPBERGER (vgl.[36],[45]) durchgeführten multivariaten Diskriminanzanalyse in der automatischen Ekg-Diagnostik eine wesentliche Reduktion der Fehldiagnosen. Die in §0 angegebenen grundlegenden Bezeichnungen entsprechen weitgehend denen von BLACKWELL und GIRSHICK [3] .

In §1 wird die Änderung des Fehlerkennungs- (=Bayes-) Risikos $s(\Gamma)$ und der Entscheidungsregionen bei Änderung der Klasseneinteilung Γ untersucht. Dabei wird, bis einschließlich §4, der Verlust jeder Fehlentscheidung gleich (mit 1) gewichtet. Vergrößert man eine (feinste) Klasseneinteilung Γ_0 mit N Klassen zu Γ , so verringert sich das Bayes-Risiko um die paarweisen Bayes-Risiken der betroffenen Klassen (I,1.7) - bis auf eine Größe $\varepsilon(\Gamma)$, die durch den Wert der jeweils drittgrößten gewichteten Dichte bestimmt wird (I,1.4 u.1.8). Da die Größen $s(\Gamma)$ und $\varepsilon(\Gamma)$ nicht exakt berechnet werden können, werden in §2 die verallgemeinerten Bhattacharyya - Koeffizienten L -ter Ordnung (BK-L) eingeführt. Mit Hilfe dieser Größen, die leicht berechnet werden können, gelingt es nun, die Bayes-Risiken nach unten und nach oben abzuschätzen (I,2.5) und den Fehlerterm $\varepsilon(\Gamma)$ nach oben zu begrenzen. Die Berechnung der BK-L für multivariate Normalverteilungen (I,2.7) zeigt nun, daß die BK-2 eine Verallgemeinerung des MAHALANOBIS - Abstands beinhalten, der bei der linearen Diskriminanzanalyse als Trennkriterium untersucht wird (I,2.8, vgl. [31]). Schließlich wird gezeigt, daß die Resultate aus den §§1 und 2 auch den Fall von Mischungen von Verteilungen (siehe [33]) erfassen (I,2.9), womit Antworten auf verschiedene Fragen gegeben werden können, die von LACHENBRUCH [26] und MICHAELIS [31] aufgeworfen worden sind (I,2.10), insbesondere die Frage nach der Übertragbarkeit von Ergebnissen des 2-Klassen-Problems auf das beliebiger Klassenanzahl und die Frage nach der Bestimmung geeigneter Verteilungsfunktionen.

Das neben dem Bayes-Risiko zweite Kriterium für die Klassenauswahl, der Informationsgehalt $\mathcal{I}$, wird in §3 eingeführt. Da für unseren Fall die Forderungen angemessen sind, die SHANNON an seine Funktion der Unsicherheit gestellt hat (vgl. ASH [2]), wird $\mathcal{I}$ hier durch das SHANNON-sche Informationsmaß dargestellt. Diese Größe ist nur von der a priori Verteilung abhängig und kann immer exakt berechnet werden. Untersucht man sie bei Klassenänderung, sieht man, daß asymptotisch der relative Informationsverlust beim Zusammenlegen der j-ten Klasse mit anderen Klassen gleich der a priori Wahrscheinlichkeit p_j ist (I,3.3).

In §4 werden zwei, aus $\mathcal{S}$ und $\mathcal{I}$ zusammengesetzte, Gütekriterien für die Klassenauswahl vorgestellt und hinreichende und notwendige Bedingungen für den Vergleich verschiedener $\mathcal{I}$ angegeben (I,4.2 u.4.5).

In §5 werden nun allgemeinere Verlustfunktionen untersucht. Dabei wird der Verlust, der bei Nichterkennen eines Musters der j-ten Klasse entsteht, mit L_j gewichtet. Es zeigt sich nun, daß dies einer Änderung der a priori Verteilung ζ zu ζ' entspricht. Im Vergleich zu §1 wird jetzt auch die Bestimmung der Risikoreduktion bei Klassenzusammenlegung komplizierter (I,5.1). Außerdem sieht man, daß sich bei Variation von L_j die Bayes-Risiken in Abhängigkeit der bedingten Risiken $\mathcal{S}(j)$ ändern (I,5.2).

Dies führt in §6 zur Diskussion des Bayes-Risikos in Abhängigkeit der a priori Verteilung und zu Aussagen über Differenzierbarkeitseigenschaften von $\mathcal{S}$ (I,6.1). Auf diesem Weg (I,6.3) erhält man auch die Charakterisierung der Minimax-Lösung, die BLACKWELL und GIRSHICK [3] gefunden haben. Außerdem wird die

Anwendbarkeit der Minimax-Lösung auf medizinische Klassifizierungsprobleme diskutiert. Da das Bayes-Risiko sich im Fall von multivariaten Normalverteilungen als differenzierbar erweist, kann man hierfür S und seine Schranken (aus I, §2) in Abhängigkeit von ζ behandeln und vergleichen (I, 6.6). Im einfachsten univariaten Fall (I, 6.8) bekommt man sogar eine Beziehung zwischen dem Bayes-Risiko, dem Informationsgehalt und den Bhattacharyya-Koeffizienten in Form einer Differentialgleichung.

Die numerische Erprobung der Resultate aus §1 bis §6 erfolgt in §7 anhand von Ekg-Daten, die von Prof. MICHAELIS geschätzt worden sind. Für diesen 4-Klassen-Fall ergibt sich, daß der Gütevergleich mit den in §4 definierten Gütekriterien bei Benutzung der S -Schranken über die BK-2 dieselben Ergebnisse erbringt wie bei Benutzung der von MICHAELIS [31] ermittelten empirischen Irrtumsraten. Dies läßt erkennen, daß in vielen Fällen die Berechnung der BK-2 das Bestimmen der empirischen Irrtumsrate durch aufwendige Testverfahren ersetzen kann. Außerdem wird (I, 7.1) der Verlauf von Irrtumsrate und Informationsgehalt bei verschiedenartiger Reduzierung der Klassenzahl am Beispiel der 7-Klassen-Diskriminanzanalyse von PIPBERGER demonstriert. Die Variation der Verlustfunktion (I, 7.2) ergibt einen etwa gleichartigen Einfluß auf J und die S -Schranken. Dies legt nahe, bei fester Klassenzahl die BK-2 als die entscheidenden Durchführungsparameter zu betrachten. Schließlich zeigt die Variation der Merkmale (I, 7.3) sehr unterschiedliche Auswirkungen auf die BK-2 und damit auf die paarweisen Risiken, was in Kap. IV, §3 zu den modifizierten Verfahren mit sequentieller Merkmalsauswahl führt.

In KAPITEL II wird die Rejectentscheidung (RE) als Bayes-Entscheidung zu einer vergrößerten Aktionsmenge eingeführt. Die hier zusätzlich gegebene Möglichkeit, schwierig zu klassifizierende Fälle zurückzuweisen, scheint mir für medizinische Verfahren besonders vorteilhaft. Gleichwohl wird sie dort in Form der RE bis jetzt noch nicht angewandt. In §1 werden zunächst das Fehlklassifizierungsrisiko E und das Reject-Risiko R definiert. Dabei zeigt es sich, daß E durch eine Größe nach oben begrenzt wird, die nur von der Wichtung des Rejectverlustes abhängt (II,1.1). Im Vergleich zur BE aus Kap.I ergibt sich, daß die RE zwar ein geringeres Fehlerkennungsrisiko, aber auch eine geringere Rate an richtigen Klassifikationen bedingt (II,1.3 u.2.2). Somit werden sowohl möglicherweise richtige wie auch falsche Bayes-Zuordnungen in einen Reject umgewandelt. Da das Fehlklassifizierungsrisiko bei der RE in beliebig kleinen Schranken gehalten werden kann, muß man die Reject-Wahrscheinlichkeit $P(V_R)$ als Kriterium für die Trennbarkeit der Klassen untersuchen. Dabei sieht man in §2, daß $P(V_R)$ zwar in ähnlicher Weise wie § paarweise zerlegbar ist, daß jedoch die Störfaktoren nur schwer zu kontrollieren sind (II,2.3).

Mit Hilfe von $P(V_R)$ kann man auch wie in I,§4 eine "Reject-Güte" der Klasseneinteilungen definieren (§3). Eine Verallgemeinerung des "Error-Reject-Tradeoffs" von CHOW [5] (II,3.1) zeigt dann, daß unter bestimmten Bedingungen die Beurteilung nach der "Reject-Güte" mit der nach der "Bayes-Güte" übereinstimmt (II,3.2). In anderen Fällen

können aber auch erhebliche Unterschiede in der Beurteilung entstehen (II,3.4).

Der Gebrauch von Abstandskriterien (KAPITEL III) ist bei der Anwendung nicht parametrischer Entscheidungsverfahren weit verbreitet (z.B. MC QUEEN[30], SEBESTYEN[39]). In Form der "Nearest-Neighbor-Rule" (NNR) haben COVER und HART [10] hauptsächlich das asymptotische Verhalten dieser Verfahren untersucht. Dabei ergab sich für das NN-Risiko eine enge Beziehung zum Bayes-Risiko ρ (III,S.3). Über eine Verknüpfung des NN-Risikos mit den paarweisen Bhattacharyya-Koeffizienten gelingt es nun hier (III,S.4), die in I,§2 angegebenen BK-Schranken für ρ zu konstruieren und Aussagen über ihre Genauigkeit zu machen. (III,S.6). Daneben zeigt es sich, daß sich die NNR bei Änderung der Klassenauswahl ähnlich wie die BE verhält (III,B.7).

Mit den in KAPITEL IV entwickelten modifizierten Entscheidungsverfahren wird versucht, die Auswahl der unterscheidbaren Klassen von dem jeweils beobachteten Muster abhängig zu machen.

Dieses Verfahren entspricht nun, in Form der modifizierten Bayes-Entscheidung (MBE) als sequentielles Verfahren durchgeführt (§1), einer gewöhnlichen Bayes-Entscheidung mit stark vergrößerter Aktionsmenge $\tilde{\mathcal{A}}$ (IV,1.2). Dabei entspricht $\tilde{\mathcal{A}}$ bei wie in I,§5 eingeschränkter Verlustfunktion der Potenzmenge der Aktionsmenge $\mathcal{A}$ aus Kap.I. Da durch die MBE gleichzeitig das Risiko jeder Entscheidung beschränkt werden kann, wird auch die RE verallgemeinert.

Die modifizierte Reject-Entscheidung (MRE), die in §2 definiert

wird, vereinfacht die Durchführung der Regel, ohne die entscheidenden Vorteile der MBE zu verlieren (IV,2.1). Eine weitere Vereinfachung bekommt man durch den Vergleich mit dem Verfahren von QUESENBERRY und GESSAMAN [37], das unter bestimmten Bedingungen mit MBE und MRE übereinstimmt (IV,2.5(2)). Dies und die Vorteile der modifizierten Regeln gegenüber BE und RE werden (IV,2.6) in einem ausführlichen Beispiel demonstriert. Dabei stellt sich heraus, daß MBE und MRE bei Beschränkung des Irrtumsrisikos einen großen Teil an Unterscheidungskraft bewahren. Der Vergleich mit der "fuzzy diagnosis" von SMETS et al.[40] stellt schließlich auch den Bezug zu aktuellen medizinischen Diagnoseverfahren her. Die zusätzlich zur MRE eingeführte sequentielle Merkmalsauswahl erfolgt auf der Basis der Ergebnisse von FU[12]. Neu ist hier der Versuch, auch die Merkmalsauswahl von bereits beobachteten Werten abhängig zu machen. Diese Vorgehensweise ist möglicherweise in der Lage, die Anzahl der zu behandelnden Merkmale zu verringern, so den Aufwand an Datenerfassung und Rechenzeit zu reduzieren und damit einige der in I,7.3 genannten Nachteile der herkömmlichen Merkmalsauswahl zu beseitigen.

Bei Verweisen innerhalb der Arbeit wird die Kapitel-Nr. nur bei Kapitelübergreifenden Hinweisen angegeben, Zur besseren Orientierung ist am Schluß der Arbeit eine Liste der verwendeten Bezeichnungen in alphabetischer Reihenfolge aufgeführt.

Ich möchte Herrn Prof. Dr. Michaelis vom Institut für medizinische Statistik und Dokumentation für die

Überlassung von Ekg-Daten zur numerischen Bearbeitung und für Hinweise auf aktuelle Literatur über medizinische Diagnoseverfahren danken.

Bei den Berechnungen, die ich am Computer des physiologischen Instituts durchführte, unterstützten mich die Mitarbeiter des Instituts für Biophysik (Leitung: Prof. Dr. Harth), Herr Wellner und Herr Dr. Brodda. Besonders hilfreich für den Fortgang der Arbeit waren dabei die Anregungen, die mir Herr Dr. Brodda in zahlreichen Diskussionen gegeben hat.

Finanziell unterstützt wurde ich von Okt. 77 bis März 78 durch ein vom Senat bewilligtes Stipendium und ab April 78 durch das Amt für Graduiertenförderung.

Schließlich danke ich Herrn Prof. Bühler, Ph.D. herzlich dafür, daß er die Betreuung dieser interdisziplinär durchgeführten Arbeit übernommen hat.

KAPITEL I

DIE BAYES ENTSCHEIDUNG

0. Generelle Voraussetzungen und Bezeichnungen

Im folgenden sei vorausgesetzt, daß jedes beobachtete Muster einen s -dimensionalen Vektor $x \in \mathbb{R}^s$ darstellt. Für $\Omega = I = \{1, \dots, N\}$, die Menge der möglichen Zustände von x , sei die a priori Verteilung $\zeta = (p_1, \dots, p_N)$ und die Menge der bedingten Dichten $\mathcal{F} = \{f_1, \dots, f_N\}$ nach einem der bekannten Verfahren geschätzt. Außerdem sei für $\mathcal{A}$, die endliche Menge der möglichen Aktionen, eine Verlustfunktion $L: \Omega \times \mathcal{A} \rightarrow \mathbb{R}$ vorgegeben und $\mathcal{X} = \{x \in \mathbb{R}^s : f(x) = \sum_{j \in I} p_j f_j(x) > 0\}$ als Merkmalsraum definiert. Weiterhin sei (vgl. [3])

(1) $\mathcal{D} = \{\varphi: \mathcal{A} \times \mathcal{X} \rightarrow [0, 1] \text{ mit } \sum_{a \in \mathcal{A}} \varphi(a|x) = 1 \quad \forall x \in \mathcal{X}\}$
die Menge der randomisierten Entscheidungsfunktionen φ ,

$$(2) \quad \mathcal{S}(\varphi|j) = \int_{\mathcal{X}} \sum_{a \in \mathcal{A}} L(j, a) \varphi(a|x) f_j(x) dx$$

das durchschnittliche Risiko von φ , gegeben $j \in \Omega$,

$$(3) \quad \mathcal{S}(\varphi) = \mathbb{E}_{\zeta}(\mathcal{S}(\varphi|\zeta)) = \sum_{j \in I} p_j \mathcal{S}(\varphi|j)$$

das durchschnittliche Risiko von $\varphi \in \mathcal{D}$,

$$(4) \quad \zeta_x(j) = (f(x))^{-1} p_j f_j(x) \text{ die a posteriori Wahrscheinlichkeit von } j, \text{ gegeben } x \in \mathcal{X},$$

$$(5) \quad \tau_x(\varphi) = \sum_{j \in I} \sum_{a \in \mathcal{A}} \zeta_x(j) L(j, a) \varphi(a|x) \text{ das a posteriori Risiko von } \varphi, \text{ gegeben } x \in \mathcal{X},$$

$$(6) \quad \mathcal{D}^* = \{\varphi^* \in \mathcal{D} : \mathcal{S}(\varphi^*) = \min_{\varphi \in \mathcal{D}} \mathcal{S}(\varphi)\} \text{ die Menge der Bayes-Entscheidungen und}$$

$$\mathcal{D}_0^* = \{\delta \in \mathcal{D}^* : \delta(a|x) \in \{0, 1\} \quad \forall a \in \mathcal{A}\} \text{ die der nicht randomisierten Bayes-Entscheidungen.}$$

Da Ω endlich ist, ist (nach [3]) $\mathcal{D}_0^* \neq \emptyset$. Für $\delta \in \mathcal{D}_0^*$ sei nun $\Omega_1 = \Omega_1(\delta) = \{x : \delta(1|x) = 1\}$ die Entscheidungs-region zu $1 \in \Omega$ (mit $\mathcal{K} = \bigcup_{i \in \Omega} \Omega_i$). $\mathcal{S} = \mathcal{S}(\delta)$ bezeichne das Bayes-Risiko.

Zunächst soll sich die Diskussion auf Bayes-Entscheidungen $\delta \in \mathcal{D}_0^*$ beschränken.

1. Änderung der Klasseneinteilung

Für jede Kategorienauswahl Γ wird das Klassifizierungsproblem durch die Größen $N(\Gamma), \Omega(\Gamma) = I(\Gamma), \mathcal{O}(\Gamma), \mathcal{T}(\Gamma), \mathcal{Z}(\Gamma)$ und $L(\Gamma)$ bestimmt. Die sich ergebenden Ausdrücke für $\mathcal{S}(\Gamma), \delta^\Gamma, \mathcal{Z}_x^\Gamma$ und τ_x^Γ lassen sich jedoch für beliebige Klasseneinteilungen i.a. nicht vergleichen. Dies liegt besonders daran, daß das Bayes-Risiko $\mathcal{S}(\Gamma)$ nur schwer und oft überhaupt nicht zu berechnen ist. So ergibt sich selbst für den 2-Klassen-Fall mit normalverteilten Dichten (mit ungleichen Kovarianz-matrizen) ein aufwendiges Verfahren (siehe [13]). Deshalb beschränken wir uns auf folgenden Fall:

Jedes Γ gehe aus einem feinsten Γ_0 durch Zusammenlegen von Klassen hervor (schreibe $\Gamma < \Gamma_0$). Damit ergibt sich mit $N = N(\Gamma_0), I = I(\Gamma_0)$ etc. und $I = \bigcup_{i=1}^M I_1, M = N(\Gamma),$

$I(\Gamma) = \{i_1, \dots, i_M\}$ mit $i_1 = \min_{j \in I_1} j,$

$p_1^\Gamma = p^\Gamma(I_1) = \sum_{k \in I_1} p_k, \quad f_1^\Gamma(x) = \sum_{k \in I_1} (p_k / p_1^\Gamma) f_k(x)$

und $\mathcal{Z}_x^\Gamma(1) = \sum_{k \in I_1} \mathcal{Z}_x(k).$

Wir schreiben dann $\Gamma = \{I_1, \dots, I_M\}.$

Nach obiger Gleichung sind die f_1^Γ Mischungen von Dichten $f_k (k \in I_1)$ mit Gewichten (p_k / p_1^Γ) . Hat man umgekehrt eine Klasseneinteilung Γ mit $f_j^\Gamma(x) = \sum_{k=1}^{N_j} q_{jk} g_{jk}(x) (\sum_{k=1}^{N_j} q_{jk} = 1$

für alle $j \in I(\Gamma)$, kann man durch $f_{jk}^{\Gamma_0} = g_{jk}$, $p_{jk}^{\Gamma_0} = p_j^{\Gamma} q_{jk}$ und $I(\Gamma_0) = \{j_k, j \in I(\Gamma), k=1, \dots, N_j\}$ ein $\Gamma_0 \succ \Gamma$ konstruieren (mit $N(\Gamma_0) = \sum_{j \in I(\Gamma)} N_j$), aus dem Γ durch Vergrößerung, d.h. Klassenzusammenlegung, hervorgeht.

Dies läßt erkennen, daß wegen der vielfältigen Auswahlmöglichkeit von Kategorien (z.B. Bildung von Subklassen) die Annahme von gemischten Verteilungen den meisten Problemen angemessen ist. Die Mischungen können dabei nach verschiedenen Methoden ([15], [16], [46]) geschätzt werden. In der Ekg - Diagnostik wurden sie bisher von Mucciardi und Gose ([33]) benutzt.

In den Paragraphen 1 - 4 dieses Kapitels soll nun der Fall $\Omega(\Gamma) = \Omega(\Gamma)$ mit $L^{\Gamma}(j, i) = 1 - \delta_{ij}$ (δ_{ij} Kronecker - Delta) behandelt werden. Allgemeinere Verlustfunktionen werden in §5 behandelt. Hier wird jetzt

$$g(\Gamma) = \sum_{\mathcal{X}} \min_{j \in I(\Gamma)} \sum_{k \in I(\Gamma), k \neq j} p_k^{\Gamma} \cdot f_k^{\Gamma}(x) dx$$

die minimale Fehlerkennungsrate. Außerdem gilt mit $\Omega_j(\Gamma) = \Omega_j(\delta^{\Gamma})$ ($\delta^{\Gamma} \in \mathcal{D}_0^*(\Gamma)$) $\{x : p_j^{\Gamma} f_j^{\Gamma}(x) > p_k^{\Gamma} f_k^{\Gamma}(x) \quad \forall k \in I(\Gamma)\} \subseteq \Omega_j(\Gamma) \subseteq \{x : p_j^{\Gamma} f_j^{\Gamma}(x) \geq p_k^{\Gamma} f_k^{\Gamma}(x) \quad \forall k \in I(\Gamma)\}$ für alle $j \in I(\Gamma)$ und alle $\delta^{\Gamma} \in \mathcal{D}_0^*$.

Für die Γ , die durch Zusammenlegen der Klassen einer Menge $J \subseteq I$ entstehen, bekommen wir

1.1 LEMMA

Sei $\Gamma = \{J, \{k\}; k \in J\}$. Dann gilt

$$g - g(\Gamma) = g(J) + \varepsilon(J) \quad \text{mit } g(J) = \sum_{j \in J} \min_{j \in J} \sum_{\substack{k \in J \\ k \neq j}} p_k f_k(x) dx$$

$$\text{und } \varepsilon(J) = \sum_{k \notin J} \int_{\mathcal{X} \setminus \bigcup_k} \left(\sum_{j \in J} p_j f_j(x) - p_k f_k(x) \right) dx$$

Dabei sind die Θ_k Entscheidungsregionen zu $\delta^\Gamma \in \mathcal{D}_0^*(\Gamma)$ mit
 $\Theta(J) = \Omega^\Gamma(J) = \{x : \sum_{j \in J} p_j f_j(x) \geq p_k f_k(x) \quad \forall k \notin J\}$.

Beweis:

$$\begin{aligned} \mathcal{S} - \mathcal{S}(\Gamma) &= \sum_{j \in J} \left(\int_{\bar{\Omega}_j} p_j f_j(x) dx - \int_{\Theta(j)} p_j f_j(x) dx \right) \\ &\quad + \sum_{k \notin J} \left(\int_{\bar{\Omega}_k} p_k f_k(x) dx - \int_{\Theta_k} p_k f_k(x) dx \right) \end{aligned}$$

(mit $\bar{\Lambda} = \mathcal{X} \setminus \Lambda$). Da aus $p_j f_j(x) \geq p_k f_k(x) \quad \forall j \in J, k \notin J$
auch $\sum_{j \in J} p_j f_j(x) \geq p_k f_k(x)$ folgt, gilt $\bar{\Omega}_j \in \Theta(J)$ und
 $\Theta_k \subseteq \bar{\Omega}_k$ für alle $j \in J, k \notin J$.

Weiter folgt mit $\Delta_J = \Theta(J) \setminus \bigcup_{j \in J} \bar{\Omega}_j = \bigcup_{j \notin J} (\bar{\Omega}_j \setminus \Theta_j)$

$$\begin{aligned} \mathcal{S} - \mathcal{S}(\Gamma) &= \sum_{j \in J} \int_{\bigcup_{\substack{k \in J \\ k \neq j}} \bar{\Omega}_k \cup \Delta_J} p_j f_j(x) dx - \sum_{k \notin J} \int_{\bar{\Omega}_k \setminus \Theta_k} p_k f_k(x) dx \\ &= \sum_{\substack{j \in J \\ \bigcup_{\substack{k \in J \\ k \neq j}} \bar{\Omega}_k}} \int p_j f_j(x) dx + \sum_{k \notin J} \int_{\bar{\Omega}_k \setminus \Theta_k} \left(\sum_{j \in J} p_j f_j(x) - p_k f_k(x) \right) dx \\ &= \int_{\bigcup_{j \in J} \bar{\Omega}_j} \min_{j \in J} \sum_{k \in J, k \neq j} p_k f_k(x) dx + \mathcal{E}(J) \quad \blacksquare \end{aligned}$$

Die neue Entscheidungsregion zu J umfaßt also zumindest alle
alten Entscheidungsregionen zu $j \in J$. Da

$$\mathcal{S}(J) = \sum_{k \in J} \int_{\bigcup_{\substack{j \in J \\ k \neq j}} \bar{\Omega}_j} p_k f_k(x) dx = \sum_{j \in J} \sum_{k \in J, k \neq j} \int_{\bar{\Omega}_k} p_j f_j(x) dx$$

ist, ergibt sich eine paarweise Darstellung von $\mathcal{S}(J)$ (vgl. [8])

1.2 FOLGERUNG

Für alle $J \in \mathcal{I}$ gilt $\mathcal{S}(J) = \sum_{\substack{j, k \in J \\ j < k}} \mathcal{S}\{j, k\}$.

1.3 DEFINITIONEN

$$a) \quad \mathcal{S}_J = \int \min_{j \in J} \sum_{k \in J, k \neq j} (p_k/p(J)) f_k(x) dx \quad (\text{für } J \subseteq I)$$

(Bayes-Risiko zur Unterscheidung der Klassen $j \in J$)

$$b) \quad \Delta(J) = p(J) \mathcal{S}_J - \mathcal{S}(J) \quad (\text{mit } p(J) = \sum_{j \in J} p_j)$$

Den Fehler in der Darstellung von Lemma 1.1 kann man abschätzen durch

1.4 LEMMA

Für alle $J \subseteq I$ gilt $0 \leq \mathcal{E}(J) \leq \Delta(J)$.

Beweis: Da für $x \in \mathcal{A}_k \setminus \bigcap_{j \in J} \mathcal{A}_j$ $\sum_{j \in J} p_j f_j(x) \geq p_k f_k(x) \geq p_1 f_1(x)$ gilt (für alle $i \in J$), folgt

$$\begin{aligned} 0 &\leq \sum_{k \notin J} \int_{\mathcal{A}_k \setminus \bigcap_{j \in J} \mathcal{A}_j} \left(\sum_{j \in J} p_j f_j(x) - p_k f_k(x) \right) dx \\ &\leq \sum_{k \notin J} \int_{\mathcal{A}_k \setminus \bigcap_{j \in J} \mathcal{A}_j} \min_{j \in J} \sum_{l \in J, l \neq j} p_l f_l(x) dx \\ &\leq \sum_{k \notin J} \int_{\mathcal{A}_k \setminus \bigcap_{j \in J} \mathcal{A}_j} \min_{j \in J} \sum_{l \in J, l \neq j} p_l f_l(x) dx = \Delta(J) \quad \square \end{aligned}$$

1.5 FOLGERUNG

Für $\Gamma = \{J, \{k\}; k \notin J\}$ gilt $\mathcal{S}(J) \leq \mathcal{S} - \mathcal{S}(\Gamma) \leq p(J) \mathcal{S}_J$.

1.6 BEMERKUNG

Man gewinnt also bei Klassenzusammenlegung höchstens das durch die Summe der entsprechenden a priori Wahrscheinlichkeiten gewichtete Risiko der betroffenen Klassen. Stellt man das Bayes-Risiko $\mathcal{S}(\Gamma_0) = \mathcal{S}$ durch paarweise Risiken dar, $\mathcal{S} = \sum_{i < j} ((p_i + p_j) \mathcal{S}_{\{i,j\}} - \Delta_{\{i,j\}})$, zeigt es sich, daß $\Delta_{\{i,j\}}$ durch den Wert der jeweils drittgrößten gewichteten Dichte bestimmt wird. Dies wird unten (siehe 1.8, 2.6) noch anhand eines Beispiels und einer Abschätzung mit Hilfe der Bhattacharyya-Koeffizienten erläutert.

1.7 SATZ

Für $M \leq N$ sei $\Gamma = \{I_1, \dots, I_M\}$. Dann gilt

$$S - S(\Gamma) = \sum_{i=1}^M S(I_i) + \varepsilon(\Gamma) \quad \text{mit} \quad \varepsilon(\Gamma) = \sum_{i=1}^M \varepsilon^{\Gamma}(I_i)$$

$$\text{und} \quad \varepsilon^{\Gamma}(I_1) = - \sum_{j \in I_1} \bigcup_{\Delta_j^{(1-1)}} \min_{j \in I_1} \sum_{k \in I_1, k \neq j} p_k f_k(x) dx \\ + \sum_{k \notin I_1} \int_{\Omega_k^{(1-1)} \setminus \Omega_k^{(1)}} \left(\sum_{j \in I_1} p_j f_j(x) - p_k^{(1-1)} f_k^{(1-1)}(x) \right) dx$$

Dabei sind $\Omega_j^{(1)}$, $p_j^{(1)}$ und $f_j^{(1)}$ die entsprechenden Größen zu $\Gamma^{(1)} = \{I_1, \dots, I_1, \{k\}; k \notin \bigcup_{j=1}^1 I_j\}$ und

$$\Delta_j^{(1)} = \Omega_j \setminus \Omega_j^{(1)} = \bigcup_{k=0}^{1-1} (\Omega_j^{(k)} \setminus \Omega_j^{(k+1)}) \quad \text{für } j \in I_{1+1}$$

Beweis: durch M-fache Anwendung von Lemma 1.1.:

Wie dort gilt beim Übergang $\Gamma^{(1-1)}$ zu $\Gamma^{(1)}$

$$\Omega_k^{(1)} \subseteq \Omega_k^{(1-1)} \quad \text{für } k \notin \bigcup_{j=1}^1 I_j$$

$$\Omega^{(1)}(I_j) \subseteq \Omega^{(1-1)}(I_j) \quad \text{für } j = 0, \dots, 1-1 \quad \text{und}$$

$$\Omega_j^{(1-1)} \subseteq \Omega^{(1)}(I_1) \quad \text{für } j \in I_1$$

Also ist $\Delta_j^{(1)}$ definiert für $j \in I_{1+1}$. Weiterhin gilt nach 1.1

$$S(\Gamma^{(1-1)}) - S(\Gamma^{(1)}) = \sum_{j \in I_1} \bigcup_{\Delta_j^{(1-1)}} \min_{j \in I_1} \sum_{\substack{k \neq j \\ k \in I_1}} p_k f_k(x) dx$$

$$+ \sum_{k \notin I_1} \int_{\Omega_k^{(1-1)} \setminus \Omega_k^{(1)}} \left(\sum_{j \in I_1} p_j f_j(x) - p_k^{(1-1)} f_k^{(1-1)}(x) \right) dx$$

$$= \sum_{\substack{j \in I_1 \\ \bigcup \Omega_j}} \min_{j \in I_1} \sum_{\substack{k \neq j \\ k \in I_1}} p_k f_k(x) dx + \sum_{k \notin I_1} \int_{\Omega_k^{(1-1)} \setminus \bigcup_{j \in I_1} \Omega_j^{(1)}} \left(\sum_{j \in I_1} p_j f_j(x) - p_k^{(1-1)} f_k^{(1-1)}(x) \right) dx$$

$$= \sum_{j \in I_1} \bigcup_{\Delta_j^{(1-1)}} \min_{j \in I_1} \sum_{k \in I_1, k \neq j} p_k f_k(x) dx = S(I_1) + \varepsilon(I_1)$$

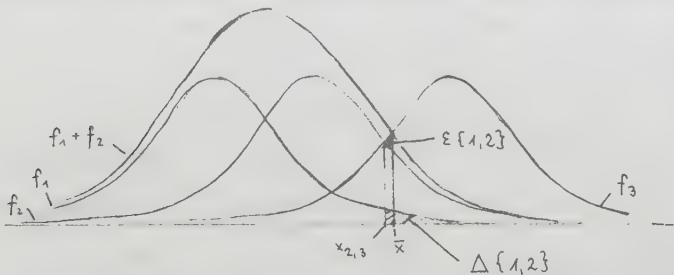
$$\begin{aligned} \text{Aus } S - S(\tau) &= \sum_{l=0}^{M-1} (S(\tau^{(1)}) - S(\tau^{(l+1)})) \\ &= \sum_{l=0}^M (S(I_l) + \varepsilon(I_l)) \text{ folgt nun die Behauptung} \end{aligned}$$

1.8 BEISPIEL

Seien $p_j = 1/N$ und $f_j \sim \mathcal{N}(\mu_j, \sigma)$ univariat normalverteilt mit $\mu_1 \leq \dots \leq \mu_N$. Dann werden die Entscheidungsgrenzen jeweils durch $(\mu_j + \mu_{j+1})/2$ markiert. Sind x_{1j} die Schnittpunkte von f_1 und f_j , ergibt sich (vgl. [45])

$$\begin{aligned} \Delta\{1,2\} &= \int_{x_{1,3}}^{\infty} p_1 f_1(x) dx = (N\sqrt{2\pi}\sigma)^{-1} \int_{(\mu_2 + \mu_1)/2}^{+\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx \\ &= (2N)^{-1} \left(1 - 2/\sqrt{\pi} \int_0^{\frac{x_{1,3} - \mu_1}{2\sigma}} e^{-x^2} dx\right). \end{aligned}$$

Dieser letzte Wert läßt sich in geeigneten Tabellen nachschlagen. Offenbar ist er umso größer (und $\Delta\{1,2\}$ umso kleiner), je größer $(x_{2,3} - \mu_1) \sigma^{-1}$ ist. Insgesamt sieht man, daß die Genauigkeit der paarweisen Darstellung von S durch Größen der Form $(\mu_j - x_{j-1,j-2}) \sigma^{-1}$ bestimmt wird.



Um die Änderung der Entscheidungsregionen zu beurteilen, muß man den Schnittpunkt $\bar{x}$ von f_3 mit $(f_1 + f_2)$ suchen. Entwickelt man dazu f_j unter Vernachlässigung aller Terme höherer als 1. Ordnung um $x_{2,3}$, bekommt man $f_3(\bar{x}) = f_1(\bar{x}) + f_2(\bar{x})$ genau

dann, wenn $\bar{x} - x_{2,3} = \frac{f_1(x_{2,3})}{f_3(x_{2,3}) - f_1(x_{2,3})}$. Auch hier werden

die Ausdrücke in Zähler und Nenner durch Größen der Form

$(\mu_j - \mu_k) \tau^{-1}$ bestimmt.

Schon für ungleiche Varianzen wird die Berechnung von $\Delta\{1,2\}$

und $\Sigma\{1,2\}$ jedoch ungleich schwieriger. Hierfür und für

allgemeinere (multivariate) Dichten sollen jetzt Schranken

entwickelt werden, die eine Abschätzung dieser Größen erlauben.

2. Abschätzungen des Bayes - Risikos

Eine obere Schranke für das Bayes-Risiko erhält man mit

(vgl. [21], [27])

2.1 SATZ

Für alle $0 \leq \alpha_{ij} \leq 1$ ($i, j \in I$) gilt

$$S \leq \sum_{\substack{i < j \\ i, j \in I}} p_i^{\alpha_{ij}} p_j^{1-\alpha_{ij}} \int_{\mathcal{X}} f_i^{\alpha_{ij}}(x) f_j^{1-\alpha_{ij}}(x) dx.$$

Beweis: Aus der Ungleichung $\min(\alpha_1, \dots, \alpha_k) \leq \alpha_1^{\gamma_1} \dots \alpha_k^{\gamma_k}$ für alle $0 \leq \alpha_1, \dots, \alpha_k$ und $0 \leq \gamma_1, \dots, \gamma_k \leq 1$ mit $\sum_{j=1}^k \gamma_j = 1$ folgt mit 1.6

$$\begin{aligned} S &\leq \sum_{i < j, i, j \in I} \int_{\mathcal{X}} \min(p_i f_i(x), p_j f_j(x)) dx \\ &\leq \sum_{\substack{i < j \\ i, j \in I}} p_i^{\alpha_{ij}} p_j^{1-\alpha_{ij}} \int_{\mathcal{X}} f_i^{\alpha_{ij}}(x) f_j^{1-\alpha_{ij}}(x) dx \quad \bullet \end{aligned}$$

Damit wird S durch die Summe leicht zu berechnender Größen, der Bhattacharyya-Koeffizienten, nach oben abgeschätzt.

2.2 DEFINITION

Für $\gamma_1, \dots, \gamma_L \geq 0$ mit $\sum_{j=1}^L \gamma_j = 1$ heißt

$$\phi(f_1, \dots, f_L; \gamma_1, \dots, \gamma_L) = \int_{\mathcal{X}} f_1^{\gamma_1}(x) \dots f_L^{\gamma_L}(x) dx$$

verallgemeinerter Bhattacharyya-Koeffizient der Ordnung L

(BK=L). Außerdem setze $\beta(f_1, \dots, f_L) = \beta(f_1, \dots, f_L; 1/L, \dots, 1/L)$

Der übliche Bhattacharyya-Koeffizient $\beta = \beta(f_1, f_2)$ liefert jetzt über das Maß von Kolmogoroff auch eine untere Schranke für das Bayes-Risiko:

2.3 SATZ

Für $N = 2$ gilt $S \geq \frac{1}{2} (1 - \sqrt{1 - 4p_1 p_2 \beta^2})$.

Beweis:

Sei $K = \frac{1}{2} \int_{\mathcal{X}} |p_1 f_1(x) - p_2 f_2(x)| dx$ das Maß von

Kolmogoroff (vgl. [23], [24]). Dann ist

$$\begin{aligned} 2K &= \int_{\Omega_1} (p_1 f_1(x) - p_2 f_2(x)) dx + \int_{\Omega_2} (p_2 f_2(x) - p_1 f_1(x)) dx \\ &= \int_{\Omega_1} p_1 f_1(x) dx + \int_{\Omega_2} p_2 f_2(x) dx - S = p_1 + p_2 - 2S, \text{ also} \end{aligned}$$

$S = 1/2 - K$. Andererseits gilt

$$\begin{aligned} 2K &= \int_{\mathcal{X}} (\sqrt{p_1 f_1(x)} - \sqrt{p_2 f_2(x)}) (\sqrt{p_1 f_1(x)} + \sqrt{p_2 f_2(x)}) dx \\ &\leq \sqrt{\int_{\mathcal{X}} (f(x) - 2\sqrt{p_1 p_2 f_1(x) f_2(x)}) dx} \sqrt{\int_{\mathcal{X}} (f(x) + 2\sqrt{p_1 p_2 f_1(x) f_2(x)}) dx} \\ &= \sqrt{(1 - 2\sqrt{p_1 p_2} \beta)(1 + 2\sqrt{p_1 p_2} \beta)} = \sqrt{1 - 4p_1 p_2 \beta^2} \text{ und damit} \end{aligned}$$

$$S \geq \frac{1}{2} (1 - \sqrt{1 - 4p_1 p_2 \beta^2})$$

Schreibt man $d_u = \sqrt{p_1 p_2} \beta$, gilt für die untere Schranke $d_1 = \frac{1}{2} (1 - \sqrt{1 - 4d_u^2})$. Also ist für zwei Paare $(\mathcal{J}_1, \mathcal{F}_1)$ und $(\mathcal{J}_2, \mathcal{F}_2)$ $d_u(\mathcal{J}_1, \mathcal{F}_1) \leq d_u(\mathcal{J}_2, \mathcal{F}_2)$ genau dann, wenn $d_1(\mathcal{J}_1, \mathcal{F}_1) \leq d_1(\mathcal{J}_2, \mathcal{F}_2)$. Auf S läßt sich diese Äquivalenz jedoch nicht ausdehnen (siehe [13]), obwohl sie an Beispielen mit empirischen Irrtumsraten zu beobachten ist (s.u.). Hierzu bewies Kailath ([23]) lediglich mit Hilfe eines Theorems

von Blackwell:

2.4 SATZ

Gegeben seien zwei Paare (f_1, f_2) und (g_1, g_2) von bedingten Dichten. Ist $\beta(f_1, f_2) \leq \beta(g_1, g_2)$, so existiert ein Paar (p_1, p_2) von a priori Wahrscheinlichkeiten, so daß auch $S(p_1, p_2, f_1, f_2) \leq S(p_1, p_2, g_1, g_2)$ ist.

Diese Eigenschaft und ähnliche, aber gröbere Abschätzungen als die über β erhält man auch für die Divergenz J von Kullback ([23],[25]):

$$J = \int_{\mathcal{X}} \ln \frac{f_1(x)}{f_2(x)} f_1(x) dx + \int_{\mathcal{X}} \ln \frac{f_2(x)}{f_1(x)} f_2(x) dx$$

Es zeigt sich jedoch ([13]), daß weder β noch J eine eindeutige oder auch nur monotone Beziehung zu S haben. Dennoch werden sie häufig bei Verfahren der Merkmalsauswahl verwendet ([45],[25]).

Mit Hilfe der "Nearest Neighbor Rule" (NNR) kann man die Abschätzungen des Bayes-Risikos auf beliebige Klassenanzahl verallgemeinern:

2.5 SATZ

Seien $f_1, \dots, f_N$ auf $\mathcal{X}$ stetig. Dann gilt mit $\beta_{ij} = \beta(f_i, f_j)$

$$\begin{aligned} \sum_{i < j} p_i p_j \beta_{ij}^2 &\leq \frac{N-1}{N} \left(1 - \sqrt{1 - \frac{2N}{N-1} \sum_{i < j} p_i p_j \beta_{ij}^2} \right) \\ &\leq S \leq \sum_{i < j} \sqrt{p_i p_j} \beta_{ij} . \end{aligned}$$

Beweis: Kap. III, F.5

Außerdem ist (siehe Kap. III und [10]) $S = 0$ (bzw. $=(N-1)/N$) genau dann, wenn $\beta_{ij}=0$ (bzw. $=1$) ist für alle $i, j \in I$.

Das von SCHWEPPE [38] vorgeschlagene Maß für die Irrtums - rate, $\beta(f_1, \dots, f_N) = \int_{\mathcal{X}} (f_1(x) \dots f_N(x))^{1/N} dx$ erfüllt dies jedoch sicher nicht und ist daher für diesen Zweck unge - eignet. Allerdings gelingt es mit den verallgemeinerten Bhattacharyya-Koeffizienten, die Fehler $\varepsilon(J)$ und $\Delta(J)$ in 1.1 und 1.3 nach oben abzuschätzen:

2.6 SATZ

Für alle $i, j \in I$ gilt

$$\varepsilon\{i, j\} \leq \Delta\{i, j\} \leq \sum_{I \ni k \neq i, j} p_i^{(\hat{i})} p_j^{(\hat{j})} p_k^{(\hat{k})} \cdot \beta(f_i, f_j, f_k; \alpha_{ijk}^{(\hat{i})}, \alpha_{ijk}^{(\hat{j})}, \alpha_{ijk}^{(\hat{k})})$$

für alle $\alpha_{ijk}^{(1)} \geq 0$ mit $\alpha_{ijk}^{(\hat{i})} + \alpha_{ijk}^{(\hat{j})} + \alpha_{ijk}^{(\hat{k})} = 1$

Beweis:

$$\begin{aligned} \Delta\{i, j\} &= \sum_{I \ni k \neq i, j} \int_{\Omega_k} \min(p_i f_i(x), p_j f_j(x)) dx \\ &= \sum_{I \ni k \neq i, j} \int_{\Omega_k} \min(p_i f_i(x), p_j f_j(x), p_k f_k(x)) dx \\ &\leq \sum_{I \ni k \neq i, j} \int_{\mathcal{X}} \min(p_i f_i(x), p_j f_j(x), p_k f_k(x)) dx \\ &\leq \sum_{I \ni k \neq i, j} p_i^{(\hat{i})} p_j^{(\hat{j})} p_k^{(\hat{k})} (f_i, f_j, f_k; \alpha_{ijk}^{(\hat{i})}, \alpha_{ijk}^{(\hat{j})}, \alpha_{ijk}^{(\hat{k})}) \end{aligned}$$

nach der in 2.1 angegebenen Ungleichung. ■

Insbesondere ergibt sich hieraus

$$0 \leq \sum_{i, j \in I, i \neq j} (p_i + p_j) S_{\{i, j\}} - S \leq \sum_{i, j \in I, i \neq j} \tilde{\beta}_{ij}$$

mit $\tilde{\beta}_{ij} = \sum_{k \neq i, j} (p_i p_j p_k)^{1/3} \beta_{ijk}$ und damit

ein Kriterium für die Genauigkeit der paarweisen Darstellung von S in 1.6. Diese Tatsache erlaubt es jetzt, zusammen mit den Abschätzungen von $S_{\{i, j\}}$, das Bayes-Risiko zu analysieren.

Damit wird es möglich, schon vor dem Testen des Entscheidungsverfahrens die Klassen anzugeben, zwischen denen die meisten Fehlklassifikationen zu erwarten sind. Bis jetzt war es stets nötig, ein aufwendiges Testverfahren durchzuführen. Außerdem können nun frühzeitig Schlüsse auf eine Änderung der Klassenauswahl gezogen werden. Dieses Problem wird in §4 noch ausführlich behandelt und in §7 anhand eines Beispiels erläutert. Zunächst soll jedoch gezeigt werden, wie die Bhattacharyya-Koeffizienten berechnet werden können.

2.7 SATZ

Seien $f_j \sim \mathcal{N}(M_j, \Sigma_j)$ s-variate Normalverteilungen mit Erwartungswert M_j und Kovarianzmatrix Σ_j (für $j=1, \dots, L$; $L \leq N$).

Dann gilt für alle $\gamma_1, \dots, \gamma_L \geq 0$ mit $\sum_{j=1}^L \gamma_j = 1$

$$\begin{aligned} \phi(f_1, \dots, f_L; \gamma_1, \dots, \gamma_L) = & \left(\prod_{j=1}^L |\Sigma_j|^{-\gamma_j/2} \right) \cdot \sqrt{|\Sigma|} \\ & \cdot \exp \left\{ -\frac{1}{2} \left(\sum_{j=1}^L \gamma_j M_j' (\gamma_j \Sigma_j^{-1}) M_j - A' \Sigma^{-1} A \right) \right\} \end{aligned}$$

mit $\Sigma = (\gamma_1 \Sigma_1^{-1} + \dots + \gamma_L \Sigma_L^{-1})^{-1}$ und

$$A = \Sigma (\gamma_1 \Sigma_1^{-1} M_1 + \dots + \gamma_L \Sigma_L^{-1} M_L), \quad |\Sigma| = \det(\Sigma),$$

falls die entsprechenden Matrizen invertierbar sind.

(A' = Transponierte von A)

Beweis:

$$(1) \quad f_1^{\gamma_1}(x) \dots f_L^{\gamma_L}(x) = \left(\prod_{j=1}^L |\Sigma_j|^{-\gamma_j/2} \right) (2\pi)^{-s/2} \cdot g(x)$$

$$\begin{aligned} \text{mit } g(x) = & \exp \left\{ -1/2 \left(\sum_{j=1}^L \gamma_j (x - M_j)' \Sigma_j^{-1} (x - M_j) \right) \right\} \\ = & \exp \left\{ -1/2 \left(\sum_{j=1}^L (x - M_j)' (\gamma_j \Sigma_j^{-1}) (x - M_j) \right) \right\} \\ = & \exp \left\{ -1/2 \left(\sum_{j=1}^L \gamma_j M_j' \Sigma_j^{-1} M_j - A' \Sigma^{-1} A \right) \right\} \\ & \cdot \exp \left\{ -1/2 ((x - A)' \Sigma^{-1} (x - A)) \right\}, \end{aligned}$$

$$\text{da } (2) \quad \sum_{j=1}^L (x - M_j)' \gamma_j \Sigma_j^{-1} (x - M_j) =$$

$$x' \Sigma^{-1} x - \sum_{j=1}^L x' \gamma_j \Sigma_j^{-1} M_j - \sum_{j=1}^L M_j' \gamma_j \Sigma_j^{-1} x + \sum_{j=1}^L M_j' \gamma_j \Sigma_j^{-1} M_j$$

$$\text{und } (x-A)' \Sigma^{-1} (x-A) = x' \Sigma^{-1} x - A' \Sigma^{-1} x - x' \Sigma^{-1} A + A' \Sigma^{-1} A$$

$$= x' \Sigma^{-1} x - \sum_{j=1}^L M_j' \gamma_j \Sigma_j^{-1} x - \sum_{j=1}^L x' \gamma_j \Sigma_j^{-1} M_j + A' \Sigma^{-1} A,$$

weil wegen $\Sigma_j = \Sigma_j'$ auch $\Sigma = \Sigma'$ und damit

$$A' = \sum_{j=1}^L \gamma_j M_j' (\Sigma_j^{-1})' \Sigma' \text{ gilt.}$$

(3) Hieraus folgt jetzt

$$\begin{aligned} \phi(f_1, \dots, f_L; \gamma_1, \dots, \gamma_L) &= \left(\prod_{j=1}^L |\Sigma_j|^{-1} \gamma_j / 2 \right) (2\pi)^{-s/2} \int_{\mathbb{R}^s} g(x) dx \\ &= \left(\prod_{j=1}^L |\Sigma_j|^{-s/2} \right) \sqrt{|\Sigma|} \exp \left\{ -1/2 \left(\sum_{j=1}^L \gamma_j M_j' \Sigma_j^{-1} M_j - A' \Sigma^{-1} A \right) \right\}. \end{aligned}$$

2.8 SPEZIALFÄLLE

(1) Für $\gamma_1 = \dots = \gamma_L = 1/L$ ergibt sich

$$\begin{aligned} \phi(f_1, \dots, f_L) &= \left(\prod_{j=1}^L |\Sigma_j|^{-1/2L} \right) \sqrt{|\bar{\Sigma}|} \\ &\quad \cdot \exp \left\{ -1/2L \left(\sum_{j=1}^L M_j' \Sigma_j^{-1} M_j - \bar{A}' \bar{\Sigma}^{-1} \bar{A} \right) \right\} \end{aligned}$$

$$\text{mit } \bar{\Sigma} = (\Sigma_1^{-1} + \dots + \Sigma_L^{-1})^{-1} = L^{-1} \Sigma$$

$$\text{und } \bar{A} = \bar{\Sigma} (\Sigma_1^{-1} M_1 + \dots + \Sigma_L^{-1} M_L) = A$$

$$\begin{aligned} (2) \text{ Hieraus folgt für } s=1 \text{ mit } \mu_j = M_j, \Sigma_j = \sigma_j^2, a_L = \sum_{j=1}^L \sigma_j^{-2}, \\ b_L = \sum_{j=1}^L (\mu_j / \sigma_j^2) \text{ und } c_L = \sum_{j=1}^L (\mu_j / \sigma_j)^2 \\ \phi(f_1, \dots, f_L) = \sqrt{L/a_L} \left(\prod_{j=1}^L \sigma_j^{-1/L} \right) e^{-1/2L(c_L - b_L^2/a_L)} \end{aligned}$$

(3) Die paarweisen Bhattacharyya-Koeffizienten lassen sich durch langwierige Matrizenumformungen vereinfachen zu

$$\phi(f_1, f_2) = |\Sigma_1|^{1/4} |\Sigma_2|^{1/4} |R/2|^{-1/2} e^{-1/4(M_1 - M_2)' R^{-1} (M_1 - M_2)}$$

mit $R = \Sigma_1 + \Sigma_2$. Für $\Sigma = \Sigma_1 = \Sigma_2$ resultiert daraus

$$\phi(f_1, f_2) = \exp((-1/2) M(f_1, f_2)). \text{ Dabei ist}$$

$M(f_1, f_2) = (M_1 - M_2)' \Sigma^{-1} (M_1 - M_2)$ der Mahalanobis - Abstand.

2.9 SATZ

Seien $g = \sum_{j \in I_1} p_j g_j$ und $h = \sum_{k \in I_2} q_k h_k$ gemischte Dichten
(mit $\sum_{j \in I_1} p_j = \sum_{k \in I_2} q_k = 1$). Dann folgt

$$\begin{aligned} 0 &\leq \sum_{j \in I_1, k \in I_2} \sqrt{p_j q_k} \beta(g_j, h_k) - \beta(g, h) \\ &\leq \sum_{\substack{(i, j) \neq (k, l) \\ i \in I_1 \times I_2}} (p_i q_j p_k q_l)^{1/4} \beta(g_i, h_j, g_k, h_l). \end{aligned}$$

Beweis: Sei $b_{jk}(x) = (p_j q_k g_j(x) h_k(x))^{1/2}$. Dann gilt

$$\begin{aligned} \sum_{(j, k) \in I_1 \times I_2} \int_{\mathcal{X}} b_{jk}(x) &= \int_{\mathcal{X}} \sqrt{\sum_{(j, k) \in I_1 \times I_2} b_{jk}^2(x) + \sum_{\substack{(i, j) \neq (k, l) \\ i \in I_1 \times I_2}} b_{ij}(x) b_{kl}(x)} dx \\ &\leq \int_{\mathcal{X}} \sqrt{\sum_{(j, k) \in I_1 \times I_2} b_{jk}^2(x) + \sum_{(i, j) \neq (k, l) \in I_1 \times I_2} b_{ij}(x) b_{kl}(x)} dx \end{aligned}$$

Dabei ist der erste Term gleich $\beta(g, h)$ und der zweite gleich

$$\sum_{(i, j) \neq (k, l) \in I_1 \times I_2} (p_i q_j p_k q_l)^{1/4} \beta(g_i, h_j, g_k, h_l) \quad \cdot$$

2.10 BEMERKUNGEN

(1) Sind zusätzlich zu den gemischten Dichten g und h aus 2.9 die a priori Wahrscheinlichkeiten p und q gegeben, kann man in gleicher Weise das Bayes-Risiko zerlegen. Konstruiert man

nämlich Γ_0 mit $I = I_1 \cup I_2$,

$$\begin{aligned} p_j^{\Gamma_0} &= p p_j \text{ für } j \in I_1 & \text{und} & & f_j^{\Gamma_0} &= g_j \text{ für } j \in I_1 \\ p_k^{\Gamma_0} &= q q_k \text{ für } k \in I_2 & & & f_k^{\Gamma_0} &= h_k \text{ für } k \in I_2 \end{aligned}$$

und legt anschließend wieder alle Klassen aus I_1 bzw. I_2

zusammen zu $\Gamma = \{I_1, I_2\}$, bekommt man $s(p, q, g, h) = s(\Gamma)$,

andererseits aber nach Satz 1.7

$$s(\Gamma_0) - s(\Gamma) = \sum_{\substack{i < j, i, j \in I_1}} s\{i, j\} + \sum_{k < l \in I_2} s\{k, l\} + \varepsilon(\Gamma).$$

Das ergibt mit $s(\Gamma_0) = \sum_{i < j \in I} s\{i, j\}$

$$\begin{aligned} s(\Gamma) &= \sum_{i \in I_1, k \in I_2} s\{i, k\} + \varepsilon(\Gamma) \\ &= \sum_{j \in I_1, k \in I_2} \int_{\mathcal{X}} \min(p_j p g_j(x), q_k q h_k(x)) dx + \varepsilon(\Gamma) = \sum_{j \in I_1, k \in I_2} \varepsilon\{j, k\}. \end{aligned}$$

Man erhält also eine zu 2.9 analoge Darstellung, deren Genauigkeit auch - nach Satz 2.6 - durch Bhattacharyya-Koeffizienten des Typs β_{ijk} bestimmt wird.

(2) Damit bekommt man jetzt auch Teillösungen der " Unsolved Practical Problems in Diskriminant Analysis", die LACHENBRUCH [26] formuliert hat:

- (i) Wie lassen sich Ergebnisse des 2-Klassen-Problems auf das endlich vieler Klassen übertragen?
- (ii) Wie ändern sich die entscheidenden Klassifizierungsgrößen, wenn man andere als Normalverteilungen betrachtet?
- (iii) Wie kann man den Fall ungleicher Kovarianzmatrizen behandeln?

Auf Frage (i) liefert die paarweise Darstellung von 1.6 eine Antwort. Dabei erweisen sich die $\Delta\{i,j\}$ als die entscheidenden Größen für die Exaktheit der Aussage. In 2.6 wird gezeigt, wie sie in einfacher Weise abgeschätzt werden können. Es zeigt sich, daß dieselben Größen auch für die Anwendung von Mischungen von Verteilungen von Bedeutung sind und so einen Hinweis zur Beantwortung von Frage (ii) geben. Was Frage (iii) betrifft, so erlauben die Anwendung der Bhattacharyya-Koeffizienten und ihre relativ leichte Berechnung für eine weite Klasse von Verteilungen, auch andere als normalverteilte Dichte und solche mit beliebigen Kovarianzmatrizen zu behandeln.

(3) Mit (1) erklärt sich auch ein Ergebnis, daß MICHAELIS [31] beobachtet hat. Er verglich zwei Verfahren zur Unterscheidung normaler Ekgs (Klasse 1) und anomaler Ekgs (Klasse 2) miteinander:

- (a) (i) Annäherung der Verteilungen f_1 und f_2 durch (multi - variate) Normalverteilungen
- (ii) Bayes - Entscheidung zum 2-Gruppen-Problem
- (b) (i) Bildung von $N-1$ Subklassen der Klasse 2
- (ii) Annäherung der jeweiligen Verteilungen durch Normal - verteilungen
- (iii) Bayes-Entscheidung δ_N zum N -Gruppen-Problem
- (iv) Bezeichnung eines Ekgs als normal, wenn es auch in (iii) als normal bezeichnet wurde, sonst als anomal

Dabei stellte MICHAELIS fest, daß das Verfahren (b) erheblich weniger Fehlklassifikationen (12,7% statt 21,6%) ergab.

Mit (1) kann jetzt eine theoretische Begründung gegeben werden: Danach entspricht (b) einer Entscheidung zum Problem (g, h) , wobei g die (normalverteilte) Dichte zu Klasse 1 und h die (Mischung von $N-1$ Normalverteilungen) zu Klasse 2 dar - stellt. Dieses h nähert aber die tatsächliche Verteilung besser an als f_2 und liefert deshalb eine geringere Fehler - kennungsrate. Bezeichnet man die Entscheidungsregionen zu (b)(iii) wieder als Ω_j , wird die Zuordnung in (b)(iv) zu

$$\varphi(x) = \begin{cases} 1, & \text{wenn } x \in \Omega_1 \\ 2, & \text{wenn } x \in \bigcup_{j=2}^N \Omega_j \end{cases} \quad . \text{ Diese kann noch ver -}$$

bessert werden durch eine Bayes-Entscheidung δ_2 . Dabei resultieren nach 1.1 die Änderungen $\Delta_{\{2, \dots, N\}}$ und $\varepsilon_{\{2, \dots, N\}}$

3. Informationsgehalt einer Klasseneinteilung

In den beiden vorangegangenen Paragraphen wurde das Verhalten der Bayes-Entscheidungen und ihrer Risiken bei verschiedener Klasseneinteilung diskutiert. Dabei wurde gezeigt, um welchen Betrag sich das Bayes-Risiko bei Klassenzusammenlegung verringert. Gleichzeitig erleidet man jedoch auch einen Verlust an Aussagekraft, wenn man auf die Unterscheidung zwischen bestimmten Klassen verzichtet. Man muß also ein Maß für die Information konstruieren, die eine gegebene Klassenauswahl, unabhängig vom jeweiligen Klassifizierungsirrtum, beinhaltet (vgl. [14]). An dieses Informationsmaß $\mathcal{J}$ wollen wir folgende Forderungen stellen:

- (1) $\mathcal{J}$ soll nur abhängig sein von der Zahl der Klassen und ihren a priori Wahrscheinlichkeiten: $\mathcal{J} = \mathcal{J}(p_1, \dots, p_N)$.
- (2) $\mathcal{J}(N) = \mathcal{J}(1/N, \dots, 1/N)$ soll mit N monoton wachsen.
- (3) $\mathcal{J}(p_1, \dots, p_N)$ soll in $p_1, \dots, p_N$ symmetrisch sein.
- (4) Teilt man die Klassifizierung in zwei Schritte, indem man zuerst nur zwischen den Klassen $\{1, \dots, r\}$ und $\{r+1, \dots, N\}$ und dann jeweils zwischen den Klassen $1, \dots, r$ und $r+1, \dots, N$ unterscheidet, so soll gelten:

$$\begin{aligned} \mathcal{J}(p_1, \dots, p_N) &= \mathcal{J}(p_1 + \dots + p_r, p_{r+1} + \dots + p_N) \\ &+ (p_1 + \dots + p_r) \cdot \mathcal{J}\left(\frac{p_1}{p_1 + \dots + p_r}, \dots, \frac{p_r}{p_1 + \dots + p_r}\right) + (p_{r+1} + \dots + p_N) \\ &\cdot \mathcal{J}\left(\frac{p_{r+1}}{p_{r+1} + \dots + p_N}, \dots, \frac{p_N}{p_{r+1} + \dots + p_N}\right). \end{aligned}$$

- (5) $\mathcal{J}(p, 1-p)$ soll in p stetig sein.

Diese Forderungen sind genau die, die SHANNON an seine Funktion der Unsicherheit H (Entropie) gestellt hat.

Nach ASH[2] ist H bis auf eine Konstante α eindeutig bestimmt:

$$\mathcal{J}(p_1, \dots, p_N) = H(p_1, \dots, p_N) = -\alpha \cdot \sum_{j \in I} p_j \ln p_j.$$

Nun ist für jedes $x \in \mathcal{X}$ und für jede Entscheidungsfunktion $\varphi \in \mathcal{D}$ $\varphi(\cdot|x)$ ein Wahrscheinlichkeitsmaß auf I . Hierfür ist die (bedingte) Entropie $H(\varphi|x) = -\alpha \sum_{j \in I} \varphi(j|x) \ln \varphi(j|x)$.

Nimmt man vor Kenntnis der bedingten Dichten eine Entscheidung

$\varphi_0(j|x) = p_j$ ($\forall j \in I$) an, so ist der Gewinn an Information durch φ $\mathcal{J}(\varphi|x, \mathcal{J}) = H(\varphi_0|x) - H(\varphi|x) = H(\mathcal{J}) - H(\varphi|x)$.

Also ist $\mathcal{J}(\varphi|x, \mathcal{J}) = H(\mathcal{J}) = \mathcal{J}(\mathcal{J})$ für jede nicht randomisierte Entscheidungsfunktion φ und damit für alle $\delta_0 \in \mathcal{D}_0^*$. Somit

ist für alle $\delta_0 \in \mathcal{D}_0^*$ $\mathcal{J}(\delta_0) = \mathcal{J}(\delta_0|x) = \mathcal{J}(\mathcal{J}) = H(\mathcal{J}) - \inf_{\varphi \in \mathcal{D}} H(\varphi|x)$

die Unsicherheit, die durch eine Entscheidung aus $\mathcal{D}$ maximal behoben werden kann.

TERRENOIRE und TOUNISSOUX [41] haben an den Informationsge-

halt $\tilde{\mathcal{J}}$ lediglich die Forderungen (1') $\tilde{\mathcal{J}}(p_1, \dots, p_N)$ ist konkav und (2') $\tilde{\mathcal{J}}(p_1, \dots, p_N) = 0$, falls ein $p_i = 1$ gestellt und ver-

schiedene Beispiele hierfür angegeben, so ein Maß ähnlich

dem von Bhattacharyya: $\tilde{\mathcal{J}}(p_1, \dots, p_N) = \sum_{i,j \in I; i \neq j} (p_i p_j)^{1/2}$. Es

zeigt sich aber, daß $\tilde{\mathcal{J}}$ die wichtige Forderung (4) nicht

erfüllt. Mit $\tilde{\mathcal{J}}$ und β bekäme man aber folgende Abschätzung:

$$\frac{1}{2} (1 - \sqrt{1 - 4 \tilde{\mathcal{J}}^2 \beta^2}) \leq \mathcal{J} \leq \tilde{\mathcal{J}} \beta.$$

Zu jeder Klassenauswahl Γ mit $N(\Gamma)$ Klassen und a priori

Verteilung $\mathcal{J}(\Gamma)$ definieren wir jetzt (mit $\alpha = 1$)

$$\mathcal{J}(\Gamma) = - \sum_{j \in I(\Gamma)} p_j^\Gamma \ln p_j^\Gamma.$$

Wir betrachten diese Größe wieder beim Vergrößern einer

Klasseneinteilung Γ_0 :

3.1 SATZ

Für $\Gamma = \{I_1, \dots, I_M\}$ ist $\mathcal{J} - \mathcal{J}(\Gamma) = - \sum_{i=1}^M \sum_{j \in I_i} p_j \ln \frac{p_j}{p(I_i)}.$

Beweis: Nach Eigenschaft (4) ist

$$\begin{aligned} J - J(\tau) &= \sum_{\ell=1}^m \left(\sum_{j \in I_\ell} p_j \right) J \left(\frac{p_{1_1}}{p(I_1)}, \dots, \frac{p_{1_m}}{p(I_1)} \right) \\ &= \sum_{\ell=1}^m \left(\sum_{j \in I_\ell} p_j \right) \left(- \sum_{j \in I_\ell} \frac{p_j}{p(I_1)} \ln \frac{p_j}{p(I_1)} \right) \\ &= - \sum_{\ell=1}^m \sum_{j \in I_\ell} p_j \ln \frac{p_j}{p(I_1)} \quad \square \end{aligned}$$

3.2 BEISPIEL

Für $p_1 = \dots = p_N = 1/N$ ist $J = J(N) = \ln N$ und $J(N) = \max_{p_1, \dots, p_N} J(p_1, \dots, p_N)$ (vgl. [2]).

Definiert man den relativen Informationsverlust beim Zusammenlegen von m Klassen, $c_N(m) = (J(N))^{-1} (J(N) - J_N(N-m+1))$, erhält man nach 3.1 (mit $M = 1, |I_1| = m$)

$$c_N(m) = \frac{m}{N} (\ln m) \frac{1}{\ln N} = \frac{m \ln m}{N \ln N}.$$

Dann ist $c_N(m) = c_N(m) - c_N(m-1)$, der Informationsverlust beim Übergang von der $(m-1)$ -ten zur m -ten Zusammenlegung, monoton wachsend in m , weil $\frac{\partial c_N}{\partial m} = \frac{1 + \ln m}{N \ln N} > 0$ und

monoton wachsend ist und daher

$$c_N(m) - c_N(m-1) > \frac{\partial c_N}{\partial m}(m-1) > c_N(m-1) - c_N(m-2) \text{ ist.}$$

Also nimmt der jeweilige zusätzliche Informationsverlust mit fortschreitender Zusammenlegung zu. Asymptotisch gilt:

Lemma $N d_N(m) \xrightarrow{m, N \rightarrow \infty} 1$, falls $\frac{\ln m}{\ln N} \xrightarrow{m, N \rightarrow \infty} 1$.

Beweis: Aus $c_N(x) = (N \ln N)^{-1} x \ln x$ folgt durch Entwickeln von c_N in $m \geq 2$:

$$N(\ln N)(c_N(x) - c_N(m)) = (1 + \ln m)(x - m) + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(n-1)! (x-m)^n}{(n+1)! m^n}$$

$$\text{und } N(\ln N)d_N(m+1) = 1 + \ln m + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(n-1)!}{(n+1)!} m^{-n}$$

da $c'(x) = (N \ln N)^{-1} (1 + \ln x)$ und

$$c^{(n)}(x) = (N \ln N)^{-1} \frac{(-1)^n (n-2)!}{x^{n-1}} \quad \text{für } n \geq 2 \text{ ist.}$$

$$\text{Weil } s(m) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(n-1)!}{(n+1)!} m^{-n} \leq s(2)$$

konvergent ist für alle $m \geq 2$, folgt

$$N d_N(m+1) = \frac{1 + \ln m}{\ln N} + \frac{s(m)}{\ln N} \quad \text{und daher}$$

$$|N d_N(m+1) - 1| \leq \left| \frac{s(m)+1}{\ln N} \right| + \left| \frac{\ln m}{\ln N} - 1 \right|$$

Beide Terme konvergieren aber für N bzw. N, m gegen ∞ gegen 0.

Die Voraussetzung $\frac{\ln m}{\ln N} \xrightarrow{m, N \rightarrow \infty} 1$ ist z.B. für alle

$m = a N$ ($1/N < a < 1$ fest) gegeben, da dann

$$\frac{\ln m}{\ln N} = \frac{\ln N - \ln(1/a)}{\ln N} \xrightarrow{N \rightarrow \infty} 1. \quad \text{Außerdem folgt dann}$$

$$d_N(m) \sim 1/N \quad \text{für } m, N \rightarrow \infty \quad (\text{d.h. } \lim_{m, N \rightarrow \infty} |d_N(m) - 1/N| = 0)$$

Das bekommt man aber auch unter schwächeren Voraussetzungen:

Satz Für alle $2 \leq m \leq N$ gilt $d_N(m) \sim 1/N$ für $N \rightarrow \infty$.

Beweis:

Sei $\varepsilon > 0$ beliebig und (nach dem Lemma)

$$N_0: 0 \leq N d_N(N) - 1 < \varepsilon \quad \forall N \geq N_0.$$

Sei jetzt $N \geq N_0$ und $N_1 = N_1(2, N)$ so gewählt, daß

$$\varepsilon_j := 1 - N d_N(j) > 0 \quad \text{für } j \leq N_1 \quad \text{und}$$

$$\varepsilon_j := N d_N(j) - 1 \geq 0 \quad \text{für } j > N_1. \quad \text{Dann gilt}$$

$$\sum_{j=2}^{N_1} \varepsilon_j = \sum_{j=N_1+1}^N \varepsilon_j < (N - N_1 - 1) \varepsilon, \quad \text{weil}$$

$$\sum_{j=2}^{N_1} d_N(j) = 1 - \sum_{j=N_1+1}^N d_N(j) \quad \text{und die } d_N(j) \text{ monoton wachsend}$$

sind. Also folgt $\varepsilon_j < (N - N_1 - 1) \varepsilon$ für alle $j = 2, \dots, N$ und

$$0 \leq |d_N(j) - 1/N| = |N d_N(j) - 1|/N \leq \varepsilon_j/N \leq (N - N_1 - 1) \varepsilon/N < \varepsilon$$

für alle $j = 2, \dots, N$.

Für beliebige a priori Verteilungen erhält man

3.3 SATZ

Gegeben sei eine Folge von a priori Verteilungen

$$\zeta(N) = (p_1^{(N)}, \dots, p_N^{(N)}) \text{ mit } p_N^{(N)} > 0.$$

Dann gilt unter den Voraussetzungen

$$(i) \lim_{N \rightarrow \infty} p_N^{(N)} = 0 \text{ und } (ii) \lim_{N \rightarrow \infty} \frac{-\ln p_N^{(N)}}{\zeta(N)} = 1$$

$$\frac{d(N)}{p_N^{(N)}} = \frac{\zeta(N) - \zeta(N-1)}{p_N^{(N)} \zeta(N)} \xrightarrow{N \rightarrow \infty} 1$$

Dabei ist $d(N)$ der relative Informationsverlust beim letzten Zusammenlegungsschritt (bzgl. $I = \{1, \dots, N\}$).

Beweis:

$$\begin{aligned} d(N) &= \zeta(p_N^{(N)}, 1-p_N^{(N)}) / \zeta(N) \\ &= (-p_N^{(N)} \ln p_N^{(N)} - (1-p_N^{(N)}) \ln (1-p_N^{(N)})) / \zeta(N) \\ &= \frac{N' \ln N' - (N'-1) \ln(N'-1)}{N' \zeta(N)} \quad \text{mit } N' = 1/p_N^{(N)}. \end{aligned}$$

Hieraus folgt jetzt

$$\begin{aligned} N' d(N) - 1 &= \frac{N' \ln N'}{\zeta(N)} - \frac{(N'-1) \ln(N'-1)}{\ln N'} \cdot \frac{\ln N'}{\zeta(N)} - 1 \\ &= \frac{N' \ln N'}{\zeta(N)} - \frac{(N'-1) \ln N' - \frac{N'-1}{x_{N'}}}{\ln N'} \cdot \frac{\ln N'}{\zeta(N)} - 1 \\ &= (N' - (N'-1)) \frac{\ln N'}{\zeta(N)} - 1 + \frac{N' - 1}{x_{N'}} \cdot \frac{1}{\zeta(N)} \\ &= \left(\frac{\ln N'}{\zeta(N)} - 1 \right) + \frac{N'-1}{x_{N'} \ln N'} \cdot \frac{\ln N'}{\zeta(N)} \end{aligned}$$

Dabei ist $x_{N'} > N'-1$ geeignet gewählt. Nun konvergiert der erste Summand gegen 0 wegen (i) und der zweite wegen (i)

und (ii). Aus $N' = 1/p_N^{(N)}$ folgt die Behauptung. ■

Auch hier gilt also auch $d(N) \sim p_N^{(N)}$ für große N . Dabei ist die Konvergenzgeschwindigkeit langsam ($-\ln p_N^{(N)}$).

3.4 SATZ

Sei $\Gamma = \{\{1, j\}, \{k\}; k \neq 1, j\}$. Dann ist

$$G(p_i, p_j) = J - J(\Gamma) = (p_i + p_j) J\left(\frac{p_i}{p_i + p_j}, \frac{p_j}{p_i + p_j}\right)$$

für festes p_i monoton wachsend in p_j (und umgekehrt).

Beweis: Mit $x = p_j$ und $a = p_i$ folgt

$$G_a(x) = G(p_i, p_j) = -x \ln\left(\frac{x}{x+a}\right) - a \ln\left(\frac{a}{x+a}\right) \text{ monoton}$$

wachsend, weil $\frac{\partial G_a(x)}{\partial x} = -\ln \frac{x}{x+a} \gg 0$ ist. ■

Für $\Gamma_m = \{\{1, \dots, m\}, \{k\}; N \geq k > m\}$ bekommt man nun wie in Beispiel 3.2 bei geordneten a priori Wahrscheinlichkeiten $p_1 \leq \dots \leq p_N$

(1) $J(\Gamma_{m-1}) - J(\Gamma_m)$ ist monoton wachsend (nach 3.4)

$$(2) J = \sum_{m=2}^N (J(\Gamma_{m-1}) - J(\Gamma_m))$$

(3) $J(\Gamma_{m-1}) - J(\Gamma_m) \leq J(\Gamma_{N-1}) - J(\Gamma_N) \sim p_N$ für große N

unter den Bedingungen von 3.3.

In diesem Fall wird somit der maximale schrittweise Informationsverlust durch die größte a priori Wahrscheinlichkeit bestimmt.

4. Gütekriterien für die Klassenauswahl

Wir haben also für jede Klassenauswahl Γ zwei Kriterien gefunden, die ihre Güte bestimmen. Dabei ist das erste Kriterium, das Bayes - Risiko, sowohl von den jeweiligen a priori Wahrscheinlichkeiten als auch von den bedingten Dichten und der Verlustfunktion abhängig und meistens nur annäherungsweise zu bestimmen. Das zweite Kriterium, der Informationsgehalt $J(\Gamma)$, ist dagegen nur von ζ abhängig und kann immer exakt berechnet werden. Beide sollen nun zu einem gemeinsamen Gütekriterium zusammengefaßt werden. Es hat verschiedene Versuche gegeben, eine (den jeweiligen Problemen) angemessene Klassenauswahl zu konstruieren. So benutzte Mc QUEEN [30] "Coarsening" und "Refinement" - Konstanten zur Konstruktion geeigneter Klasseneinteilungen. SEBESTYEN verwendete (in [39]) ein Verfahren ähnlich der "Nearest Neighbor Rule" (vgl. Kap III und [10]), das das Eröffnen neuer Klassen ermöglichte. Für die automatische Ekg - Diagnostik diskutierten u.a. PIPBERGER [36] und MICHAELIS [31] die Bildung von Subklassen und zeigten auf, daß in praxi oft die Anzahl der zur Verfügung stehenden Fälle hierfür begrenzend ist.

Für alle diese Verfahren sind jedoch keine Beziehungen zu den entscheidenden Klassifikationsgrößen, der Fehlklassifikationsrate und der gewünschten Aussagekraft, untersucht worden. Von diesem Entscheidungs- und Informations-orientierten Standpunkt aus werden jetzt zwei Kriterien für die Güte einer Klassenauswahl vorgestellt:

Für den Vergleich beliebiger Klasseneinteilungen führen wir

das Quotientenkriterium ein.

4.1 DEFINITION

Γ_1 heit (echt) Q-besser als Γ_2 bei minimalem Informations -
gehalt γ , falls

$$\frac{g(\Gamma_1)}{g(\Gamma_2)} \leq \frac{j(\Gamma_1)}{j(\Gamma_2)} \quad \text{fr } j(\Gamma_1), j(\Gamma_2) \geq \gamma.$$

Γ heit Q-optimal, wenn es keine echt Q-bessere Klassenaus-
wahl gibt.

Sind (fr $L(j,i)=1-\delta_{ij}$) d_u und d_l die oberen bzw. unteren
Schranken des Bayes-Risikos nach Satz 2.5, erhlt man

4.2 SATZ

$$d_u(\Gamma_1) j(\Gamma_2) \leq d_l(\Gamma_2) j(\Gamma_1) \quad \text{ist hinreichend und}$$

$$d_l(\Gamma_1) j(\Gamma_2) \leq d_u(\Gamma_2) j(\Gamma_1) \quad \text{ist notwendig dafr,}$$

da Γ_1 Q-besser ist als Γ_2 .

Werden nur solche Klasseneinteilungen miteinander verglichen,
die durch Zusammenlegen von Klassen aus einem gemeinsamen
feinsten Γ_0 hervorgehen, untersucht man das Differenzen -
kriterium:

4.3 DEFINITION

Seien $\Gamma_1, \Gamma_2 \prec \Gamma_0$. Dann heit Γ_1 (echt) D-besser als Γ_2
(schreibe $\Gamma_1 \succ \Gamma_2$), wenn

$$\frac{g(\Gamma_0) - g(\Gamma_1)}{g(\Gamma_0)} - \frac{j(\Gamma_0) - j(\Gamma_1)}{j(\Gamma_0)} > \frac{g(\Gamma_0) - g(\Gamma_2)}{g(\Gamma_0)} - \frac{j(\Gamma_0) - j(\Gamma_2)}{j(\Gamma_0)}$$

Γ heit D-optimal, wenn es keine echt D-bessere Klassen -
auswahl gibt.

4.4 BEMERKUNGEN

(1) Setzt man $c(\Gamma) = J^{-1}(J - J(\Gamma))$ (mit $S = S(\Gamma_0)$, $J = J(\Gamma_0)$), so ist Γ_1 D-besser als Γ_2 genau dann, wenn

$$(S - S(\Gamma_1)) - c(\Gamma_1) \cdot S \geq (S - S(\Gamma_2)) - c(\Gamma_2) \cdot S \iff$$

$$S(\Gamma_1) + c(\Gamma_1) \cdot S \leq S(\Gamma_2) + c(\Gamma_2) \cdot S \quad \text{ist.}$$

Dabei kann $c(\Gamma)S$, wie bei sequentiellen Entscheidungsverfahren (siehe [12], [3] und Kap.IV) als die entstandenden Kosten (hier: an Informationsgehalt) beim Übergang $\Gamma_0 \rightarrow \Gamma$ aufgefaßt werden. Diese Interpretation führt nun zu modifizierten Bayes-Entscheidungen, die in Kap.IV behandelt werden.

(2) Da $S - S(\Gamma) - S c(\Gamma) = \left(\frac{S}{J} - \frac{S(\Gamma)}{J(\Gamma)} \right) J(\Gamma)$ ist (für $J(\Gamma) > 0$), ist Γ genau dann D-besser als Γ_0 , wenn es Q-besser ist als Γ_0 . Beim Vergleich zweier $\Gamma_1, \Gamma_2 < \Gamma_0$ zeigt obige Gleichung aber auch, daß bei Q-gleich guten Γ_1 und Γ_2 das Differenzenkriterium die Klasseneinteilung mit größerem Informationsgehalt bevorzugt. Das ist in den meisten Fällen eine gewünschte Erscheinung.

Bemerkt sei noch, daß wegen der Quotientenbildung in 4.1 und 4.3 die Forderung $\alpha = 1$ in §3 keine Einschränkung darstellt.

(3) Wie in 4.2 lassen sich auch Bedingungen für die D-Reihenfolge der Klasseneinteilungen angeben. Dabei sind die Abschätzungsverluste (nach 1.7 und 2.5) für nur wenig verringerte Klassenanzahl kleiner als in 4.2. Da sie im wesentlichen durch die Abschätzung von $S(\Gamma_0)$ zustandekommen, sind sie weiter zu verringern, wenn man in 4.3 statt mit $S(\Gamma_0)$ und $J(\Gamma_0)$ mit den maximal möglichen Werten, 1 und $\ln N$, normiert. Man untersucht dann $S - S(\Gamma) - (\ln N)^{-1}(J - J(\Gamma))$.

4.5 BEISPIEL

Sei $N(\Gamma) = N-1$. Beim Zusammenlegen der Klassen i und j ($\Gamma = \{\{i, j\}, \{k\}; k \neq i, j\}$) bekommt man unter Vernachlässigung von $\{i, j\}$ nach Lemma 1.1

$$S - S(\Gamma) = (\ln N)^{-1} (J - J(\Gamma)) = (p_i + p_j) (s_{ij} - J(q_{ij}, q_{ji}) / \ln N)$$

mit $s_{ij} = S_{\{i, j\}}$ und $q_{ij} = p_i / (p_i + p_j)$.

Dann gilt $S_{ij} - J(q_{ij}, q_{ji}) / \ln N \leq 0$ und damit $\Gamma \stackrel{D}{<} \Gamma_0$,

$$\text{wenn } q_{ij}^\alpha q_{ji}^{1-\alpha} \beta(f_i, f_j; \alpha) = (\ln N)^{-1} J(q_{ij}, q_{ji}) \leq 0$$

für ein $0 \leq \alpha \leq 1$ ist. Für $\alpha = q_{ij}$ bekommt man damit als hinreichende Bedingung für $\Gamma_0 \stackrel{D}{>} \Gamma$

$$\ln N \beta(f_i, f_j; q_{ij}) \leq J(q_{ij}, q_{ji}) \exp(J(q_{ij}, q_{ji})).$$

Nach 2.8 wird hieraus für $p_i = p_j$ und $f_k \sim N(\mu_k, \sigma_k^2)$ ($k = i, j$)

$$\ln N \left(\frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} + 2 \ln \left(\frac{\sigma_i^2 + \sigma_j^2}{2 \sigma_i \sigma_j} \right) \right) \leq 8 \ln 2.$$

4.6 BEMERKUNG

In allen gebräuchlichen Ekg-Diagnoseverfahren ist es üblich, für jeden Patienten nur das Vorliegen einer von N Krankheiten anzunehmen. Dies ist i.a. der Realität nicht angemessen. HABBEMA [17] hat deshalb versucht, eine Klasseneinteilung mit allen 2^N Krankheitskombinationen zu behandeln. Da das wegen der begrenzten Fallzahl und des großen Rechenaufwands Schwierigkeiten machte, entwickelte er zwei Kriterien, die bei einer N -Klasseneinteilung einen Hinweis auf das Vorliegen mehrerer Krankheiten gaben. Da eines davon die Überlappung der bedingten Dichten betrifft, sind zur Erkennung solcher Fälle auch die verallgemeinerten Bhattacharyya-Koeffizienten geeignet. Dieses Thema wird bei der Behandlung der modifizierten Verfahren in Kap. IV noch einmal aufgegriffen.

5. Einflüsse der Verlustfunktion

Da eine Ausdehnung der Ergebnisse von §1 - §4 auf beliebige Verlustfunktionen erhebliche Schwierigkeiten bereitet, beschränken wir uns hier auf solche folgenden Typs:

$$(L1) \quad L^{\Gamma}(J_1, J_2) = L^{\Gamma}(J_1) \text{ für } J_1 \neq J_2 \text{ und } = 0 \text{ für } J_1 = J_2 \text{ für} \\ \text{alle } J_1, J_2 \in \Gamma.$$

$$(L2) \quad L(J) = L^{\Gamma_1}(J) = L^{\Gamma_2}(J) \text{ für alle } J \in \Gamma_1, \Gamma_2 \text{ und } \Gamma_1, \Gamma_2 \text{ beliebig}$$

$$(L3) \quad L_j = L\{j\} \leq L(J) \text{ für alle } j \in J \leq I.$$

Somit ist das Nichterkennen eines Musters der Klasse $j \in J$ nur von J abhängig, gleichgültig in welche (falsche) Klasse es eingeordnet wird. Dabei wird die Größe von $L(J)$ durch Faktoren wie Prognose, Therapierbarkeit und Ansteckungsgefahr der Krankheiten $j \in J$ bestimmt. Obwohl in den meisten medizinischen Diagnoseverfahren die Betrachtung von Verlustfunktionen dieses Typs ausreichend ist, sei bemerkt, daß man sich mit dieser Einschränkung die Möglichkeit nimmt, Fehlklassifikationen mit geringen Folgen (etwa die Verwechslung von Erkrankungen mit ähnlicher Therapie und Prognose) gering zu wichten.

Setzt man jetzt für $\Gamma = \{I_1, \dots, I_M\}$ und $j \in I$

$$p_j^{\Gamma} = p_j L(J) / \bar{L}^{\Gamma} \quad \text{mit} \quad \bar{L}^{\Gamma} = \sum_{J \in \Gamma} p(J) L(J)$$

für alle $j \in J \in \Gamma$, bekommt man für $\Gamma = \Gamma_0$

$$f(x|\mathcal{J}) \tau_x(i|\mathcal{J}, L) = \sum_{j \neq i, j \in I} p_j L_j f_j(x) = \bar{L} \sum_{j \neq i} p_j^{\Gamma} f_j(x) \\ = \bar{L} f(x|\mathcal{J}') \tau_x(i|\mathcal{J}') \quad \text{und} \quad g(\mathcal{J}, L) = \bar{L} \cdot g(\mathcal{J}').$$

Durch Ersatz von $\mathcal{J}$ durch $\mathcal{J}'$ kann man nun alle Abschätzungen von g aus §2 auf den Fall (L1)-(L3) übertragen.

5.1 SATZ

Sei $\Gamma = \Gamma_J = \{J, \{k\}; k \notin J\}$. Dann gilt

$$S(J') - S^\Gamma(J', \Gamma) = \sum_{j \in J} \Omega_j(J') \min_{k \neq j, k \in J} \sum_{k \in J} p_k' f_k(x) dx + \varepsilon(J, L).$$

Dabei ist $\varepsilon(J, L) = \varepsilon(J, J') + \varepsilon_0(J, L)$ mit

$$\varepsilon_0(J, L) = \sum_{k \neq j} \int_{\bar{\Theta}_k'} (p_k' - p_k'^\Gamma) f_k(x) dx - \int_{\bar{\Theta}_j'} \sum_{j \in J} (p_j'^\Gamma - p_j') f_j(x) dx$$

Außerdem gelte wie in 1.1 $\bar{\Theta}_k' \subseteq \Omega_k'$ für $k \notin J$ und $\Omega_j' \subseteq \bar{\Theta}_j'$ für $j \in J$.

Beweis: Nach (L1) - (L3) sind die Entscheidungsregionen in der angegebenen Art zu konstruieren, weil aus

$$p_j L_j f_j(x) > p_k L_k f_k(x) \text{ für alle } j \in J, k \notin J \text{ auch}$$

$$L(J) \sum_{j \in J} p_j f_j(x) > p_k L_k f_k(x) \text{ folgt. Aus}$$

$$S(J') - S(J', \Gamma) = \sum_{j \in I} \int_{\bar{\Omega}_j'} p_j' f_j(x) dx - \sum_{k \neq j} \int_{\bar{\Theta}_k'} p_k'^\Gamma f_k(x) dx$$

$$- \int_{\bar{\Theta}_j'} \sum_{j \in J} p_j'^\Gamma f_j(x) dx$$

$$= \sum_{j \in I} \int_{\bar{\Omega}_j'} p_j' f_j(x) dx - \sum_{k \neq j} \int_{\bar{\Theta}_k'} p_k' f_k(x) dx - \int_{\bar{\Theta}_j'} \sum_{j \in J} p_j' f_j(x) dx$$

+ $\varepsilon_0(J, L)$ folgt dann die Behauptung nach Lemma 1.1. ■

Offenbar ist die Abbildung $J' = J'(L)$ für jedes feste J

mit $p_j > 0 \quad \forall j \in I$ und

$$p_j' = \begin{cases} p_j L_j / \bar{L} & \text{für } L \neq (0, \dots, 0) \\ 0 & \text{für } L = (0, \dots, 0) \end{cases}$$

eine Abbildung von R^N in $\mathcal{Q} = \{q \in [0, 1]^N; \sum_{j \in I} q_j = 1\}$. Außerdem ist sie surjektiv, da für jedes $q \in \mathcal{Q}$ mit $L_j = q_j / p_j$ $q = J'(L)$ gilt. Also ist jede Änderung der Verlustfunktion vom Typ

(L1)-(L3) äquivalent zur Änderung der a priori Wahrscheinlichkeiten. Verändert man auf diese Weise nur eine a priori Wahrscheinlichkeit, bekommt man

5.2 LEMMA

Seien $\zeta = (p_1, \dots, p_N)$ und $\zeta' = (q_1, \dots, q_N)$

mit $q_j = p_j L_j / \bar{L}$ für $L_j = \begin{cases} L+1 & \text{für } j = 1 \\ 1 & \text{für } j \geq 2 \end{cases}$ mit $L > 0$.

Dann folgt

$$\begin{aligned} \sum_{j=2}^N (p_j - q_j) (s(\delta_Y | 1) - s(\delta_Y | j)) &\geq s(\zeta') - s(\zeta) \\ &\geq \sum_{j=2}^N (p_j - q_j) (s(\delta_{Y'} | 1) - s(\delta_{Y'} | j)) \end{aligned}$$

Dabei sind δ_Y und $\delta_{Y'}$ Bayes-Entscheidungen zu ζ bzw. ζ' , für die $\delta_Y(x) = j \iff \delta_{Y'}(x) = j$ für alle $x \in \Omega_{jk} = \{x: p_j f_j(x) = p_k f_k(x)\} = \{x: q_j f_j(x) = q_k f_k(x)\} \quad \forall j, k \geq 2$ gilt.

Beweis: Mit $\vartheta_j = p_j - q_j$ ($j \geq 2$) und $\vartheta = q_1 - p_1$ gilt

$$\begin{aligned} s(\zeta') - s(\zeta) &= \sum_{j \in I} \int_{\bar{\Omega}_j'} q_j f_j(x) dx - \sum_{j \in I} \int_{\bar{\Omega}_j} p_j f_j(x) dx \\ &= \int_{\bar{\Omega}_1'} p_1 f_1(x) dx - \int_{\bar{\Omega}_1} p_1 f_1(x) dx + \vartheta \int_{\bar{\Omega}_1'} f_1(x) dx \\ &\quad + \sum_{j \geq 2} \left(\int_{\bar{\Omega}_j'} p_j f_j(x) dx - \int_{\bar{\Omega}_j} p_j f_j(x) dx - \vartheta_j \int_{\bar{\Omega}_j'} f_j(x) dx \right) \\ &= \vartheta \int_{\bar{\Omega}_1'} f_1(x) dx - \sum_{j \geq 2} \vartheta_j \int_{\bar{\Omega}_j'} f_j(x) dx + \sum_{j \geq 2} \left(\int_{\bar{\Omega}_j \setminus \bar{\Omega}_j'} (p_j f_j(x) - p_1 f_1(x)) dx \right) \end{aligned}$$

Dies gilt, da für $x \in \bar{\Omega}_1$ aus $p_1 f_1(x) \geq p_j f_j(x)$ auch $q_1 f_1(x) > q_j f_j(x)$ folgt (weil $q_1 > p_1, q_j < p_j$ und $f_1(x) > 0$) und da nach Konstruktion von δ_Y und $\delta_{Y'}$ analog $\bar{\Omega}_j' \subseteq \bar{\Omega}_j$ ist. Also sind die Ausdrücke $\bar{\Omega}_j \setminus \bar{\Omega}_j'$ erklärt (mit $\bar{\Omega}_1' \setminus \bar{\Omega}_1 = \bigcup_{j \geq 2} \bar{\Omega}_j \setminus \bar{\Omega}_j'$).

Außerdem hat man für $x \in \Omega_j \setminus \Omega_j'$

$$0 \leq p_j f_j(x) - p_1 f_1(x) = (p_j - \delta_j) f_j(x) - (p_1 + \delta) f_1(x) + \delta_j f_j(x) + \delta f_1(x) \leq \delta f_1(x) + \delta_j f_j(x) \quad \text{und daher}$$

$$\begin{aligned} g(\zeta') - g(\zeta) &\geq \delta \int_{\Omega_1'} f_1(x) dx - \sum_{j \geq 2} \delta_j \int_{\Omega_j'} f_j(x) dx \\ &= \sum_{j \geq 2} \delta_j \left(\int_{\Omega_1'} f_1(x) dx - \int_{\Omega_j'} f_j(x) dx \right) \quad \text{und} \end{aligned}$$

$$\begin{aligned} g(\zeta') - g(\zeta) &\leq \delta \int_{\Omega_1'} f_1(x) dx + \sum_{j \geq 2} \int_{\Omega_j \setminus \Omega_j'} \delta f_1(x) dx \\ &\quad - \sum_{j \geq 2} \left(\delta_j \int_{\Omega_j'} f_j(x) dx - \delta_j \int_{\Omega_j \setminus \Omega_j'} f_j(x) dx \right) \\ &= \delta \int_{\Omega_1'} f_1(x) dx - \sum_{j \geq 2} \delta_j \int_{\Omega_j'} f_j(x) dx = \sum_{j \geq 2} \delta_j \left(\int_{\Omega_1'} f_1(x) dx - \int_{\Omega_j'} f_j(x) dx \right) \end{aligned}$$

Das Bayes-Risiko ändert sich also umso stärker, je weiter die entsprechenden bedingten Fehlerkennungswahrscheinlichkeiten auseinanderliegen. Dabei ist, weil

$$0 \leq g(\delta_{\zeta} | j) = \int_{\Omega_j} f_j(x) dx \leq 1 \quad \text{ist,} \quad |g(\zeta') - g(\zeta)| \leq q_1 - p_1.$$

Da für $N=2$ und $p_1 < q_1, p_2 > q_2$ sicher $\Omega_1(\zeta) \subseteq \Omega_1(\zeta')$ und $\Omega_2(\zeta') \subseteq \Omega_2(\zeta)$ ist, kann man wie in 5.2 beweisen:

5.3 FOLGERUNG

Sei $N=2$, $\zeta = (p_1, p_2)$, $\zeta' = (q_1, q_2)$ und $p_1 < q_1$. Dann gilt

$$\begin{aligned} (q_1 - p_1) (g(\delta_{\zeta} | 1) - g(\delta_{\zeta} | 2)) &\geq g(\zeta') - g(\zeta) \\ &\geq (q_1 - p_1) (g(\delta_{\zeta'} | 1) - g(\delta_{\zeta'} | 2)) \end{aligned}$$

für alle Bayes-Entscheidungen δ_{ζ} bzw. $\delta_{\zeta'}$ zu ζ bzw. ζ' .

6. Einflüsse der a priori Verteilungen

In §5 wurde gezeigt, daß eine Änderung der Verlustfunktion eines bestimmten Typs einer Änderung der a priori Wahrscheinlichkeiten entspricht. Da in den meisten Entscheidungsverfahren in der Medizin die a priori Verteilung nur schwer zu bestimmen ist, ist die Untersuchung ihres Einflusses auf die Entscheidungsregionen und -risiken von Bedeutung.

Wir setzen nun (für $L(j,i)=1-\delta_{ij}$) $g(\gamma)$ von $\mathcal{A}$ auf $[0,1]^N$

fort durch
$$g(\gamma) = \int_{\mathcal{X}} \min_{j \in I} \sum_{k \neq j} p_k f_k(x) dx .$$

$\Omega_k(\gamma)$ sei wieder so gewählt, daß $\mathcal{X} = \bigcup_{k \in I} \Omega_k(\gamma)$ und $\{x: p_k f_k(x) > p_j f_j(x) \ \forall j \neq k\} \subseteq \Omega_k(\gamma) \subseteq \{x: p_k f_k(x) \geq p_j f_j(x) \ \forall j \neq k\}$ für alle $\gamma \in [0,1]^N$ gilt. Setzt man für $A \subseteq \mathcal{X}$

$P_k(A) = P(A|k) = \int_A f_k(x) dx$, bekommt man

6.1 SATZ

Vor.: (V) Für $k \in I$ und $\gamma \in (0,1)^N$ konvergiere $P_k(\Omega_k(\gamma + \vartheta_k) \setminus \Omega_k(\gamma))$ und $P_k(\Omega_k(\gamma) \cap \Omega_k(\gamma - \vartheta_k))$ für ϑ gegen 0^+ gegen 0 (mit $\vartheta_k = (0, \dots, \vartheta, \dots, 0)$) für geeignete $\Omega_j(\gamma)$, $\Omega_j(\gamma - \vartheta_k)$, $\Omega_j(\gamma + \vartheta_k)$ mit $\Omega_j(\gamma + \vartheta_k) \subseteq \Omega_j(\gamma) \subseteq \Omega_j(\gamma - \vartheta_k)$ für $j \neq k$.

Beh.: g ist in γ partiell nach p_k differenzierbar und

$$\frac{\partial g}{\partial p_k}(\gamma) = P_k(\bar{\Omega}_k).$$

Beweis: Sei o.B.d.A. $k=1$. Dann folgt für $\vartheta > 0$

$$\begin{aligned} g(\gamma + \vartheta_1) - g(\gamma) &= \sum_{j \neq 1} \int_{\bar{\Omega}_j(\gamma + \vartheta_1) \setminus \bar{\Omega}_j(\gamma)} p_j f_j(x) dx - \int_{\bar{\Omega}_1(\gamma) \setminus \bar{\Omega}_1(\gamma + \vartheta_1)} p_1 f_1(x) dx \\ &\quad + \vartheta \int_{\bar{\Omega}_1(\gamma + \vartheta_1)} f_1(x) dx \\ &= \sum_{j \neq 1} \int_{\Omega_j(\gamma) \setminus \Omega_j(\gamma + \vartheta_1)} (p_j f_j(x) - p_1 f_1(x)) dx + \vartheta \int_{\bar{\Omega}_1(\gamma + \vartheta_1)} f_1(x) dx. \end{aligned}$$

Für $x \in \Omega_j(\gamma) \setminus \Omega_j(\gamma + \vartheta_1)$ ($j \neq 1$) folgt jetzt

$$0 \leq p_j f_j(x) - p_1 f_1(x) = p_j f_j(x) - (p_1 + \vartheta) f_1(x) + \vartheta f_1(x) \leq \vartheta f_1(x), \text{ also}$$

$$0 \leq (\vartheta(\gamma + \vartheta_1) - \vartheta(\gamma)) - \vartheta \int_{\bar{\Omega}_1(\gamma + \vartheta_1)} f_1(x) dx \leq \vartheta \int_{\Omega_1(\gamma + \vartheta_1) \setminus \Omega_1(\gamma)} f_1(x) dx$$

$$\text{und, da } \int_{\bar{\Omega}_1(\gamma)} f_1(x) dx = \int_{\bar{\Omega}_1(\gamma + \vartheta_1)} f_1(x) dx + \int_{\Omega_1(\gamma + \vartheta_1) \setminus \Omega_1(\gamma)} f_1(x) dx,$$

$$0 \leq \left| \frac{\vartheta(\gamma + \vartheta_1) - \vartheta(\gamma)}{\vartheta} - \int_{\bar{\Omega}_1(\gamma)} f_1(x) dx \right| \leq \int_{\Omega_1(\gamma + \vartheta_1) \setminus \Omega_1(\gamma)} f_1(x) dx.$$

Der letzte Ausdruck konvergiert aber nach (V) gegen 0 für ϑ gegen 0^+ . Analog folgt die Behauptung für ϑ gegen 0^- . ■

Man sieht mit diesem Beweis auch, daß auch ohne die Voraussetzung (V) die Stetigkeit von $\vartheta(\gamma)$ auf $(0,1)^N$ folgt. Durch Einschränkung auf $\mathcal{Q}$ erhält man damit (vgl. BLACKWELL - GIRSHICK[3]):

6.2 SATZ

$\vartheta(\gamma)$ ist stetig auf $\mathcal{Q}$.

6.3 SATZ

Unter der Voraussetzung (V) aus 6.1 für alle $k \in I$ gilt:

Jedes lokale Extremum γ^0 auf $\mathcal{Q}$ erfüllt

(E) $\vartheta(\delta_{\gamma^0} | j) = \vartheta(\delta_{\gamma^0} | k)$ für alle $k, j \in I$ und eine geeignete Bayes-Entscheidung δ_{γ^0} zu γ^0 .

Beweis: Sei $g(\gamma) = \sum_{j \in I} p_j - 1$ für $\gamma \in [0,1]^N$.

Dann ist $\frac{\partial g}{\partial p_j} = 1$ für alle $j \in I$. Da außerdem $\frac{\partial \vartheta}{\partial p_j} = P_j(\bar{\alpha}_j)$

ist nach 6.1, gilt nach der Methode von Lagrange für jedes

lokale Extremum auf $\mathcal{Q} = [0,1]^N \cap \{\gamma : g(\gamma) = 0\}$

$\vartheta(\delta_{\gamma^0} | j) + \lambda = 1$ (λ Lagrange-Multiplikator) für alle $j \in I$.

Ist $g(\gamma^1 | \gamma) = \sum_{j \in I} p_j g(\delta_{\gamma^1 | j})$ das Risiko, wenn man eine

Bayes-Entscheidung zu γ^1 statt zu γ wählt, so gilt mit

$$\alpha(\gamma^0) = g(\delta_{\gamma^0} | j) \quad g(\gamma^0 | \gamma) = \sum_{j \in I} p_j \alpha(\gamma^0) = \alpha(\gamma^0) \quad \forall \gamma \in Q$$

Weil für beliebiges $\gamma^1 \in Q$

$g(\gamma^1 | \gamma^0) \geq g(\gamma^0) = \alpha(\gamma^0)$ gilt, ist γ^0 Minimax-Lösung,

$$\text{d.h. } \sup_{\gamma \in Q} g(\gamma^0 | \gamma) = \min_{\gamma^1 \in Q} \sup_{\gamma \in Q} g(\gamma^1 | \gamma).$$

6.4 SATZ

Ist (V) aus 6.1 für alle $j \in I$ erfüllt, so nimmt $g(\gamma)$

sein Maximum auf Q in γ^0 an und γ^0 erfüllt die Bedingung

(E) aus 6.3.

Beweis: Da g stetig ist, muß es auf der kompakten Menge sein Maximum annehmen. Nach 6.3 genügt es zu zeigen, daß γ^0 nicht auf dem Rand ∂Q von Q liegt. Dazu sei

$$\gamma = (0, p_2, \dots, p_N) \in \partial Q.$$

(1) Da $f(x) > 0$ ist für alle $x \in X$, ist $\Omega_1 = \Omega_1(\gamma) = \emptyset$,

$$\text{also } P_1(\bar{\Omega}_1) = 1.$$

Sei $p_1 = \max_{i \in I} p_i$. Weil $\int_{\Omega_k} f_1(x) dx > 0$ für ein $k \neq 1$ und

für $x \in \Omega_k$ $f_k(x) \geq (p_1/p_k) f_1(x) \geq f_1(x)$ gilt, ist

$$\int_{\Omega_k} f_k(x) dx \geq \int_{\Omega_k} f_1(x) dx > 0 \text{ für ein } 1 \neq k \in I, \text{ also}$$

$$a_k := P_1(\bar{\Omega}_1) - P_k(\bar{\Omega}_k) > 0.$$

(2) Sei jetzt γ' so gewählt, daß $p_1' > 0$ beliebig und

$p_j' = p_j(1-p_1')$ für alle $j \neq 1$ gilt. Dann geht γ aus γ' ge-

mäß Lemma 5.2 hervor (mit $L=-1$). Wählt man wie dort Ω_j, Ω_j' ,

so gilt (γ_j wie im Beweis von 5.2) $\sum_{j \geq 2} \gamma_j \left(\int_{\bar{\Omega}_1} f_1(x) dx - \int_{\bar{\Omega}_j} f_j(x) dx \right)$

$$\geq g(\gamma') - g(\gamma) \geq \sum_{j \geq 2} \gamma_j' \left(\int_{\bar{\Omega}_1'} f_1(x) dx - \int_{\bar{\Omega}_j'} f_j(x) dx \right), \text{ also}$$

$$\begin{aligned}
 s(\gamma') - s(\gamma) &= \sum_{j \geq 2} \gamma_j \left(\int_{\bar{A}_j \setminus \bar{A}_j'} f_1(x) dx + \int_{\bar{A}_j' \setminus \bar{A}_j} f_j(x) dx \right) \\
 &= p_1' (a_k p_k - \sum_{j \geq 2} p_j \left(\int_{\bar{A}_j \setminus \bar{A}_j'} f_1(x) dx + \int_{\bar{A}_j' \setminus \bar{A}_j} f_j(x) dx \right))
 \end{aligned}$$

Ist $0 < \varepsilon < a_k p_k$, so gibt es nach (V) ein γ' , so daß

$$0 \leq \sum_{j \geq 2} p_j \left(\int_{\bar{A}_j \setminus \bar{A}_j'} f_1(x) dx + \int_{\bar{A}_j' \setminus \bar{A}_j} f_j(x) dx \right) < \varepsilon \quad \text{und damit} \quad .$$

$s(\gamma') - s(\gamma) > p_1' (a_k p_k - \varepsilon) > 0$, also $s(\gamma') > s(\gamma)$ gilt. ■

Außerdem hat man für alle γ , die für ein $\delta_\gamma(E)$ aus 6.3 erfüllen: $s(\gamma) = s(\delta_\gamma | \gamma^0) > s(\gamma^0) = s(\delta_\gamma | \gamma) > s(\gamma)$, also $s(\gamma) = s(\gamma^0)$. Damit ist jede a priori Verteilung, die (E) genügt, "least favorable" (d.h. $s(\gamma) = \max_{\gamma \in \mathcal{Q}} s(\gamma^1)$). Umgekehrt ist nach WALD[44] jede Minimaxlösung eine Bayes-Entscheidung zu γ^0 mit (E). Also stellt $\{\gamma: \gamma \text{ erfüllt (E)}\}$ die Menge aller "least favorable distributions" dar und bestimmt die Menge der Minimaxlösungen.

Andererseits zeigen die Betrachtungen in §5 und am Anfang dieses Paragraphen, daß für die medizinischen Diagnoseverfahren Minimaxlösungen nicht geeignet sind. Hiermit vergibt man nämlich die Möglichkeit der Gewichtung der Fehlklassifizierungen. Außerdem kann man den Einfluß der unterschiedlichen Häufigkeiten der verschiedenen Krankheiten nicht mehr berücksichtigen. Dies hat auch schon MICHAELIS [31] festgestellt. Allerdings zeigt es sich, daß ein Fehler in der Annahme der a priori Verteilung dann nur geringe Folgen hat, wenn das (durch die Wahl der Verlustfunktion) resultierende γ' nahe an der "least favorable distribution" liegt. Genauer werden die Auswirkungen durch 5.2 beschrieben.

Für $N=2$ ergeben sich wieder einfachere Verhältnisse:

6.5 SATZ

Vor.: Sei $N=2$ und $p=p_j$. Für $p \in (0,1)$ konvergiere $P_j(\Omega_1(p') \setminus \Omega_1(p))$ gegen 0 für p' gegen p^+ bei geeigneten $\Omega_1(p)$ (bzw. $P_j(\Omega_2(p') \setminus \Omega_2(p))$ gegen 0 für p' gegen p^-) für $j=1,2$.

Beh.: $\mathcal{S}$ ist rechtsseitig (linksseitig) differentierbar in p

$$\text{und} \quad \frac{\partial \mathcal{S}}{\partial p^+} = \mathcal{S}(\mathcal{S}_p | 1) - \mathcal{S}(\mathcal{S}_p | 2) \quad (= \frac{\partial \mathcal{S}}{\partial p^-})$$

für eine geeignete Bayes-Entscheidung $\mathcal{S}_p$.

Beweis: Nach 5.3 gilt

$$\int_{\bar{\Omega}_1} f_1(x) dx - \int_{\bar{\Omega}_1} f_2(x) dx \geq \frac{\mathcal{S}(p') - \mathcal{S}(p)}{p' - p} \geq \int_{\bar{\Omega}_1} f_1(x) dx - \int_{\bar{\Omega}_2} f_2(x) dx,$$

$$\text{also } 0 \leq \left| \frac{\mathcal{S}(p') - \mathcal{S}(p)}{p' - p} \right| - \left(\int_{\bar{\Omega}_1} f_1(x) dx - \int_{\bar{\Omega}_2} f_2(x) dx \right)$$

$$= \int_{\bar{\Omega}_1} f_1(x) dx - \int_{\bar{\Omega}_2} f_2(x) dx - \frac{\mathcal{S}(p') - \mathcal{S}(p)}{p' - p} \leq \int_{\bar{\Omega}_1 \setminus \bar{\Omega}_1'} (f_1(x) + f_2(x)) dx.$$

Nun konvergiert der letzte Ausdruck gegen 0 für p' gegen p^+ .

6.6 BEMERKUNG

Für das Bayes-Risiko erhält man somit im Fall $N=2$

$$-1 \leq \frac{\partial \mathcal{S}}{\partial p} \leq 1, \quad \frac{\partial \mathcal{S}}{\partial p} \text{ monoton fallend, } \frac{\partial \mathcal{S}}{\partial p} \xrightarrow{p \rightarrow 0^+} 1,$$

$$\frac{\partial \mathcal{S}}{\partial p} \xrightarrow{p \rightarrow 0^-} -1 \text{ und } \frac{\partial \mathcal{S}}{\partial p} = 0, \text{ falls } \mathcal{S}(\mathcal{S}_p | 1) = \mathcal{S}(\mathcal{S}_p | 2) \text{ ist.}$$

Dagegen gilt für die Schranken des Bayes-Risikos

$$\frac{\partial d}{\partial p} u^- = \frac{1-2p}{2(p(1-p))^{1/2}} \neq 0, \quad \frac{\partial^2 d}{\partial p^2} u^- \leq 0 \text{ und}$$

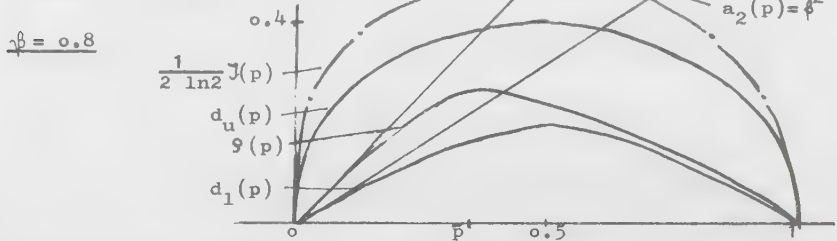
$$\frac{\partial d}{\partial p} 1^- = \frac{p^2(1-2p)}{(1-4p(1-p))^{1/2}} \neq 0, \quad \frac{\partial^2 d}{\partial p^2} 1^- \leq 0.$$

Beide sind also konkave Funktionen mit Maximum in $1/2$ und

$$\frac{\partial d}{\partial p} u^-(0) = - \frac{\partial d}{\partial p} u^-(1) = +\infty, \quad \frac{\partial d}{\partial p} l^-(0) = - \frac{\partial d}{\partial p} l^-(1) = \beta^2.$$

Da $\frac{\partial J}{\partial p}(p) = \ln((1-p)/p)$ und $\frac{\partial J}{\partial p}(0) = - \frac{\partial J}{\partial p}(1) = +\infty$ ist,

bekommt man folgendes Diagramm



Hier hat es den Anschein, daß (für diesen Verlauf von g , mit $\bar{p} < 1/2$) für $p < \bar{p}$ gut durch $a_1(p)=p$ und für $p > \bar{p}$ gut durch $d_l(p)$ angenähert werden kann. Dies geht umso besser, je größer β ist. Außerdem hat man die Möglichkeit, durch Variation von β in $d_u(\beta) = p^\beta(1-p)^{1-\beta} \varphi(\beta)$ die Abschätzungen zu verbessern. Dieses Problem wird in III, 8.7 noch einmal aufgegriffen.

6.7 BEISPIEL

Man sieht jetzt, daß die Voraussetzung des Satzes 6.5 für normalverteilte Dichten und für alle $0 < p_1, p_2 < 1$ erfüllt ist. Es gilt nämlich für $p_1' > p_1$

$\Omega_1(p_1') \setminus \Omega_1(p_1) \subseteq \{x: p_1' f_1(x) > p_2 f_2(x) > p_1 f_1(x)\}$. Also erfüllen für $f_j \sim \mathcal{N}(M_j, \Sigma_j)$ alle $x \in \Omega_1(p_1') \setminus \Omega_1(p_1)$

$$0 \leq \ln(p_2/\sqrt{|\Sigma_2|}) - \ln(p_1/\sqrt{|\Sigma_1|}) + \frac{1}{2} (x-M_1)' \Sigma_1^{-1} (x-M_1) - \frac{1}{2} (x-M_2)' \Sigma_2^{-1} (x-M_2) \leq \ln(p_1'/p_1).$$

a) $\Sigma_1 = \Sigma_2 = \Sigma$. Hierfür wird diese Ungleichung zu

$A \leq x' \Sigma^{-1} (M_2 - M_1) + (M_2 - M_1)' \Sigma^{-1} x \leq A + \varepsilon(p_1', p_1)$ (A geeignet).

Die Lösungen stellen einen s -dimensionalen Bereich zwischen Hyperebenen der Form $\{x: \sum_{j=1}^s a_j x_j = a\}$ und $\{x: \sum_{j=1}^s a_j x_j = a + \varepsilon\}$ dar. Durch geeignete maßerhaltende Transformationen kann man sehen, daß dieser Bereich ein Maß $\lambda^s(\varepsilon(p_1', p_1))$ besitzt, das für p_1' gegen p_1 gegen 0 konvergiert.

Für $s=1$ wird $\lambda(\mathcal{L}_1(p_1') \setminus \mathcal{L}_1(p))$ durch die Lösungsmenge der Ungleichung $0 \leq A + (x(\mu_2 - \mu_1) + b) / (2\sigma^2) \leq \ln(p_1' / p_1)$ bestimmt, deren Maß offensichtlich gegen 0 für p_1' gegen p_1 strebt.

b) $\mathcal{L}_1 \neq \mathcal{L}_2$.

Hier muß man im eindimensionalen Fall Lösungen der quadratischen Ungleichung

$$0 \leq A + \frac{1}{2} (x^2 (\sigma_1^{-2} - \sigma_2^{-2}) + x(\mu_2 \sigma_2^{-2} - \mu_1 \sigma_1^{-2}) + \mu_1^2 - \mu_2^2) \leq \varepsilon$$

betrachten ($\varepsilon = \ln(p_1' / p_1)$). Für beliebige Dimension wird die Lösungsmenge durch s -dimensionale Ellipsoide bestimmt (siehe ANDERSON [1]). Eine geometrische Veranschaulichung wurde auch von MICHAELIS [31] erarbeitet.

Für den einfachsten Fall ($s=1, \sigma_1 = \sigma_2 = \sigma$) erhält man eine einfache Beziehung zwischen dem Bayes-Risiko und dem Informationsgehalt in Form einer Differentialgleichung:

6.8 SATZ

Seien $f_j \sim \mathcal{N}(\mu_j, \sigma)$ für $j=1,2$. Dann gilt:

$\vartheta(p)$ ist zweimal nach p differenzierbar auf $(0,1)$ und

$$\frac{\partial^2 \vartheta}{\partial p^2} = \frac{\sigma^2}{\mu_1 - \mu_2} \frac{\partial^2 \vartheta}{\partial p^2}(p) (f_1 + f_2)(x_0) \text{ mit } x_0 = \frac{\sigma^2}{\mu_1 - \mu_2} \frac{\partial \vartheta}{\partial p}(p) + \frac{\mu_1 + \mu_2}{2}.$$

Beweis: Sei $g(x) = f_1(x) / f_2(x)$ der Likelihood-Quotient und

$q = (1-p)/p$. Dann ist

$g(x) = \exp\left(\frac{\mu_1 - \mu_2}{\sigma^2} - (x - (\mu_1 + \mu_2)/2)\right)$ und

$$g^{-1}(x) = \frac{\sigma^2 \ln x}{\mu_1 - \mu_2} + \frac{\mu_1 + \mu_2}{2} \quad (\text{für } x \neq 0) \quad \text{mit } (g^{-1})'(x) = \frac{\sigma^2}{(\mu_1 - \mu_2)x}.$$

Seien jetzt $\nu_1, \nu > 0$ so, daß $q - \nu = (p_2 - \nu_1)/(p_1 + \nu_1)$ ist.

$$\text{Dann ist } \frac{\partial g}{\partial p}(p + \nu_1) - \frac{\partial g}{\partial p}(p) = - \int_{\Omega_1(p + \nu_1) \setminus \Omega_1(p)} (f_1(x) + f_2(x)) dx.$$

Dabei sind die Entscheidungsregionen so gewählt, daß

$$\Omega_1(p + \nu_1) \setminus \Omega_1(p) = \{x: q - \nu < g(x) \leq q\} = g^{-1}(q - \nu, q] \text{ ist.}$$

Nimmt man o.B.d.A. $\mu_1 > \mu_2$ an, ist g monoton wachsend und

$$\text{daher } \lambda \{g^{-1}(q - \nu, q]\} = g^{-1}(q) - g^{-1}(q - \nu) = (g^{-1})'(\tilde{q}(\nu)) \cdot \nu$$

für ein $q - \nu \leq \tilde{q}(\nu) \leq q$. Hieraus folgt nun

$$\begin{aligned} & \frac{\nu}{\nu_1} (g^{-1})'(\tilde{q}(\nu)) \sup_{x \in g^{-1}(q - \nu, q]} (f_1(x) + f_2(x)) > \\ & - \frac{\partial g(p + \nu_1)/\partial p - \partial g(p)/\partial p}{\nu_1} > \frac{\nu}{\nu_1} (g^{-1})'(\tilde{q}(\nu)) \inf_{x \in g^{-1}(q - \nu, q]} (f_1(x) + f_2(x)) \end{aligned}$$

Da $\nu = \nu_1/(p_1^2 + \nu_1)$, konvergiert $(\partial g/\partial p(p + \nu_1) - \partial g/\partial p(p))/\nu_1$ gegen $p^{-2}(g^{-1})'(q)(f_1 + f_2)(g^{-1}(q))$.

Ein analoger Beweis für ν gegen 0 ergibt dann

$$\frac{\partial^2 g}{\partial p^2}(p) = - \frac{1}{p^2} (g^{-1})'(q) (f_1 + f_2)(g^{-1}(q)).$$

$$\text{Für } f_j \sim \mathcal{N}(\mu_j, \sigma^2) \text{ folgt dann mit } \frac{\partial^2 g}{\partial p^2}(p) = - \frac{1}{p(1-p)}$$

$$\frac{\partial^2 g}{\partial p^2}(p) = - \frac{1}{p^2} \frac{\sigma^2}{\mu_1 - \mu_2} \frac{p}{1-p} (f_1 + f_2)(x_0).$$

Für $p=1/2$ bekommt man eine Faktorisierung von $\partial^2 g/\partial p^2$:

$$\frac{\partial^2 g}{\partial p^2}(1/2) = \frac{\partial^2 g}{\partial p^2}(1/2) \frac{2}{\sqrt{2\pi}\sigma}.$$

Wie man an dem Beweis des Satzes erkennt, kann man die zweite Ableitung des Bayes-Risikos auch für allgemeinere Dichten

berechnen. Dabei stellt sich heraus, daß das Verhalten von f_1 und f_2 auf der Entscheidungsgrenze $S = \{x: g(x) = q\}$ von Bedeutung ist. Für den univariaten Fall ist dann die Abzählbarkeit von S zusammen mit der lokalen Invertierbarkeit des Likelihood-Quotienten g und einigen lokalen Differenzierbarkeitsvoraussetzungen von g^{-1} für die Existenz der zweiten Ableitung hinreichend.

7. Numerische Ergebnisse

Als Beispiel behandelte ich die automatische Ekg - Analyse von vier kongenitalen Herzvitien: Pulmonalstenose (Klasse 1), Vorhofseptumdefekt(2), Fallot'sche Tetralogie(3), Ventrikel - septumdefekt(4).[†] Für jede der vier Klassen wurde eine (11-variate) Normalverteilung angenommen. Die geschätzten Eigenwerte und Kovarianzmatrizen wurden mir von Prof. MICHAELIS zur Verfügung gestellt. Mit Hilfe eines Fortran IV - Computer Programms berechnete ich

- (i) die Bhattacharyya-Koeffizienten BK-2 und BK-3
- (ii) die oberen und unteren Schranken für das Gesamtrisiko θ und für die partiellen Risiken θ_{ij} .
- (iii) die entsprechenden Informationsgrößen $\mathcal{I}$ und $\mathcal{I}_{ij}$ (Definition s.u.).

7.1 ANALYSE DER KLASSENEINTEILUNGEN

Für vier ausgewählte Ekg - Merkmale ergab sich im Vergleich mit den entsprechenden von MICHAELIS [31] gefundenen empirischen Irrtumsraten $\tilde{\theta}$ und $\tilde{\theta}_{ij}$:

$$d_1(\theta) = 0.067 \leq \tilde{\theta} = 0.19 \leq d_u(\theta) = 0.515,$$

$$\text{also } d_1(\theta) = 0.35\tilde{\theta} \text{ und } \tilde{\theta} = 0.37d_u(\theta).$$

[†] Im Vergleich zu [31] sind die Klassennummern 1 und 2 vertauscht.

Beim Vergleich der paarweisen (partiellen) Risiken erhielt man eine monotone Beziehung zwischen den über die BK-2 ermittelten Schranken und den empirischen Risiken:

$$d_1(g_{ij}) \leq d_1(g_{kl}) \Leftrightarrow \tilde{g}_{ij} \leq \tilde{g}_{kl} \Leftrightarrow d_u(g_{ij}) \leq d_u(g_{kl}).$$

Die Reihenfolge war $\tilde{g}_{14} \succ \tilde{g}_{12} \succ \tilde{g}_{13} \succ \tilde{g}_{24} \succ \tilde{g}_{34} \succ \tilde{g}_{23}$.

Die zugehörigen Tabellen sind im einzelnen in [20] aufgeführt.

Die sehr unterschiedlich großen paarweisen Risiken

($d_u(g_{14}) = 7 d_u(g_{23})$) sind dabei für den großen Wert der Varianz des a posteriori Risikos (vgl. III, 5.6 und III, B.7) und damit für die schlechte Annäherung von g durch $d_1(g)$ verantwortlich.

Die Berechnung der Bhattacharyya-Koeffizienten 3. Ordnung ergab, gemessen an der monotonen Beziehung zwischen $d_u(g_{ij})$ und $\tilde{g}_{ij}$, große Werte der ϕ_{ijk} :

$$\phi_{2,3,4} = 0.10 \leq \phi_{1,2,3} = 0.15 \leq \phi_{1,3,4} = 0.18 \leq \phi_{1,2,4} = 0.29.$$

Mit der Bestimmung von $\phi'_{ijk} = \min(\phi_{ij}, \phi_{ik}, \phi_{jk})$ und

$\phi''_{ijk} = \phi_{ij} \phi_{ik} \phi_{jk}$ (nach der Hölderschen Ungleichung gilt $\phi_{ijk} \leq (\phi'_{ijk})^{2/3} \leq (\phi''_{ijk})^{2/9}$) erhält man dieselbe Reihenfolge:

$$\phi'_{2,3,4} = 0.162 \leq \phi'_{1,2,3} = 0.162 \leq \phi'_{1,3,4} = 0.201 \leq \phi'_{1,2,4} = 0.250$$

$$\phi''_{2,3,4} = 0.008 \leq \phi''_{1,2,3} = 0.018 \leq \phi''_{1,3,4} = 0.056 \leq \phi''_{1,2,4} = 0.061.$$

Nach der Analyse des Bayes-Risikos mit Hilfe der Bhattacharyya-Koeffizienten bzw. der empirischen Irrtumsraten sind die Klassen 1 und 4 diejenigen, die sich am meisten, und die Klassen 2 und 3 diejenigen, die sich am wenigsten überlappen. Vergleicht man danach bei schrittweisem Zusammenlegen verschiedene Klasseneinteilungen nach dem Differenzenkriterium

$D(r) = (d_u(s) - d_u(s(r))) / d_u(s) = (J - J(r)) / J$ und nach

dem Quotientenkriterium $Q(r) = d_u(s(r)) / J(r)$ für

a) $r_1^a = \{\{1,4\}, \{3\}, \{2\}\}$, $r_2^a = \{\{1,2,4\}, \{3\}\}$ und

b) $r_1^b = \{\{2,3\}, \{1\}, \{4\}\}$, $r_2^b = \{\{2,3,4\}, \{1\}\}$,

bekommt man mit fallender D- und Q-Güte

r	r_1^a	r_2^a	r_0	r_1^b	r_2^b
$D(r)$	0.10	0.05	0	-0.14	-0.29
$Q(r)$	0.32	0.32	0.38	0.44	0.61

Bemerkenswert ist, daß trotz $Q(r_1^a) = Q(r_2^a)$ $D(r_1^a) = 2 D(r_2^a)$

ist, daß also, wie in 4.4(2) betont, die Klassenauswahl mit größerem Informationsgehalt vom D-Kriterium deutlich bevorzugt wird.

Diese Betrachtungen machen es nun möglich, schon vor dem Bestimmen der empirischen Irrtumsraten durch aufwendige Testverfahren, die Risiken in ihrer Größe und Zusammensetzung zu beurteilen und Schlußfolgerungen auf eine geeignete Klassenauswahl zu ziehen.

Bemerkung: Ich habe die Änderung des Fehlerkennungsrisikos und des Informationsgehalts auch am Beispiel der linearen Diskriminanzanalyse (Bayes-Entscheidung zu Normalverteilungen mit gemeinsamer "gepoolter" Kovarianzmatrix) von 7 Klassen durchgeführt. Die Kovarianzmatrix und die Eigenwerte (66 Merkmale) sind von MICHAELIS [31] mit von Prof. PIPBERGER ermittelten Vektor-Ekg-Daten geschätzt worden. Für die Annäherung des Bayes-Risikos benutzte ich hier die empirischen (siehe [31]) Irrtumsraten bei der Annahme gleicher a priori Wahrscheinlichkeiten.¹⁾

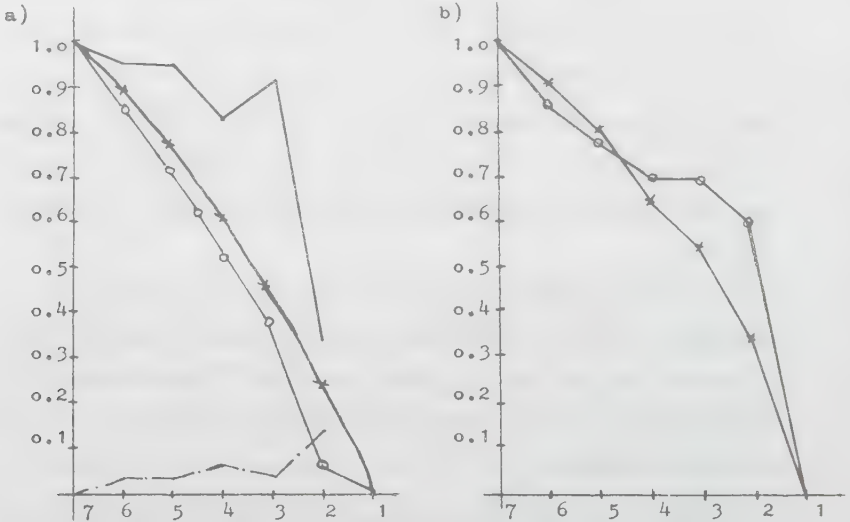
¹⁾ Die Berechnung der Bhattacharyya-Koeffizienten für diesen Fall ist in Bearbeitung.

Für die Zusammenlegungsreihenfolgen

a) (1,5); (1,2,5); (1,2,5,7); (1,2,3,5,7); (1,2,3,5,6,7)

b) (1,5); (1,5), (2,7); (1,5), (2,3,7); (1,5), (2,3,7), (4,6);
(1,4,5,6), (2,3,7)

bekam ich folgende Diagramme



Es ist die empirische Irrtumsrate $\frac{\tilde{\gamma}(\Gamma)}{\gamma}$ (○—○) und der Informationsgehalt $\frac{\gamma(\Gamma)}{\gamma}$ (×—×) gegen die jeweils verbleibende Klassenanzahl aufgetragen. In a) ist zusätzlich $\frac{Q(\Gamma)}{\gamma}$ (—) und $D(\Gamma)$ eingezeichnet.

Man sieht, daß im Fall a) alle Zerlegungen besser sind als Γ_0 . Dabei wird für $N=2$ die beste Zusammenlegung erreicht, wieder mit besonders starker Abnahme des Gütequotienten α (da N klein). Da dann jedoch nur noch $1/4$ des ursprünglichen Informationsgehalts übrig ist, sollte diese Möglichkeit durch die Forderung $\gamma(\Gamma) > \alpha$ (wie in §4) verhindert werden. Während in a) sowohl $\tilde{\gamma}(\Gamma)/\gamma$ als auch $\gamma(\Gamma)/\gamma$ fast linear (und parallel)

abnehmen (bei jedem Schritt etwa um $1/6$, vgl. 3.2 und 3.3), fallen diese Größen im Fall b) erst bei den letzten Zusammenlegungsschritten ab. Dabei bleibt mehr als 60% des ursprünglichen Risikos bis zum letzten Schritt erhalten (7% in a). Andererseits ist in b) bei 3 Klassen noch über die Hälfte an Information vorhanden ($1/3$ in a). Dies liegt daran, daß in b) die a priori Wahrscheinlichkeiten sich weniger unterscheiden (also $I(\Gamma)$ groß) und daß hier nicht so viele paarweise Risiken beim Zusammenlegen entfallen wie in a) (nach Satz 1.7).

7.2 EINFLUSS VON VERLUSTFUNKTION UND A PRIORI VERTEILUNG

Als Beispiel behandelte ich eine Wichtung der Klasse 3 (die Fallotsche Tetralogie ist das schwerste Herzvitium) der kongenitalen Herzvitien mit $L_3=5$ bzw. $L_3=10$. Mit den resultierenden a priori Verteilungen γ' (nach §5) erhielt ich

L_3	1	5	10
$d_u(s)$	0.515	0.409	0.323
I	1.363	1.210	0.962
$d_u(s)/I$	0.38	0.34	0.34

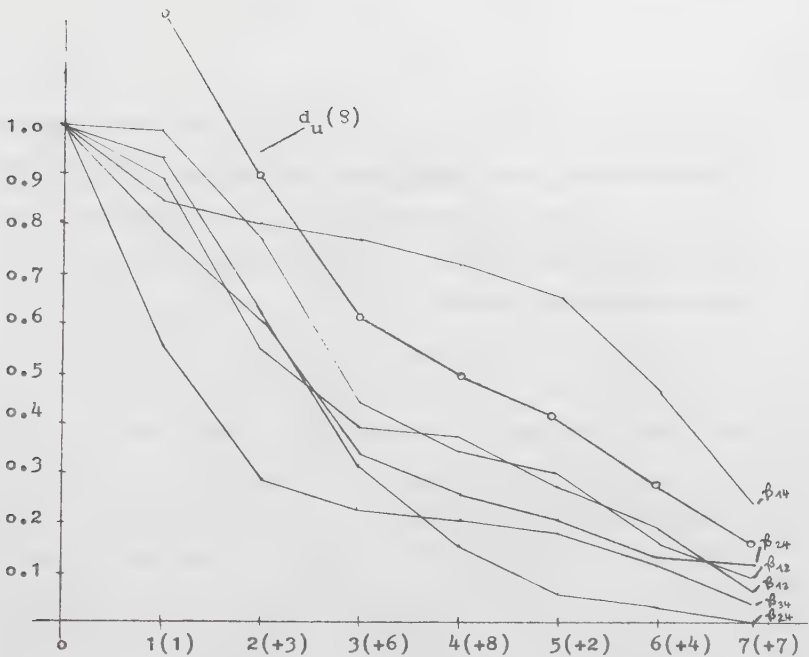
Es resultieren somit für $L_3=5$ und $L_3=10$ etwas Q-bessere Klasseneinteilungen. Auch die Reihenfolge der paarweisen Risiken und Informationen ($I_{ij} = -p_i \ln(p_i/(p_i+p_j)) - p_j \ln(p_j/(p_i+p_j))$) ist der Informationsverlust beim Zusammenlegen der Klassen i und j) änderte sich. Dabei blieben $d_u(s_{ij})$ und I_{ij} für $i=3$ oder $j=3$ ungefähr gleich, während sich die übrigen partiellen Größen stark verringerten, allerdings in s und I in gleicher Weise. Genau ergab sich:

$j \backslash i$	1	2	3	1	2	3
2	2.62			2.62		
3	0.83	0.83		0.98	1.03	
4	2.62	2.62	0.83	2.62	2.62	0.95
	$d_u(s_{ij}; L_3=1)/d_u(s_{ij}; L_3=10)$			$J_{ij}(L_3=1)/J_{ij}(L_3=10)$		

Die etwas größeren Werte rechts bewirken den etwas günstigeren Güte-Quotienten. Ähnliche Ergebnisse bekommt man auch für $L_3=5$. Eine Wichtung von Klasse 1 ist in [20] erläutert. Da einerseits $\tilde{s}(3)=0.07 < \tilde{s}(1)=0.13 < \tilde{s}(2)=0.28 < \tilde{s}(4)=0.78$ ($\tilde{s}(j)$ bedingte empirische Risiken) ist, andererseits eine zusätzliche Wichtung von Klasse 3 $q(3)$ weiter verkleinert (aus $L_j^\uparrow$ folgt $\Omega_j^\uparrow$ und $q(j) \downarrow$), wird sich die resultierende a priori Verteilung weiter von der "least favorable distribution" entfernen und so immer ein geringeres Gesamtrisiko ergeben. Dies kann man auch bei der 7-Klassen-Diskriminanzanalyse von PIPBERGER (vgl. MICHAELIS [31]) finden. Dabei läßt der gleichartige Einfluß einer Änderung der Verlustfunktion auf Bayes-Risiko und Informationsgehalt es sinnvoll erscheinen, bei fester Klassenanzahl die Bhattacharyya-Koeffizienten als die entscheidenden Durchführungsparameter zu betrachten.

7.3 EINFLUSS DER MERKMALSAUSWAHL

Berechnet man die paarweisen Bhattacharyya-Koeffizienten und die Schranken für das Bayes-Risiko jeweils bei successiver Hinzunahme neuer Merkmale, bekommt man folgenden Verlauf:



$d_u(s)$ und β_{1j} sind gegen die Merkmalsanzahl (in Klammern die jeweils hinzugenommene Merkmals-Nr. nach [31]) abgetragen.

Während die Abnahme von $d_u(s)$ gleichmäßig erfolgt, ist die der BK-2 sehr unterschiedlich. Dies zeigt die gegenseitige Abhängigkeit von geeigneter Merkmals- und Klassenauswahl. Außerdem zeigen sich schon für 7 Merkmale unterschiedliche Ergebnisse bei alternativer Berechnung der BK-2 nach 2.7 und 2.8(3). Diese fortgepflanzten Rundungsfehler sind wegen der größeren Anzahl aufwendiger Matrizeninversionen und Determinantenberechnungen bei den BK-3 noch eher zu erwarten. Zwar kann dieser Fehler durch die oben (7.1) angegebene Abschätzung der BK-3 durch Funktionen der BK-2 weitgehend reduziert werden, doch wird man für eine große Merkmalsanzahl, wie

etwa bei den PIPBERGER-Daten, mit sehr ungenauen Ergebnissen rechnen müssen. Aus diesem Grund wird in Kap.IV ein Verfahren vorgestellt, das die Merkmalsauswahl flexibel handhabt und so die Merkmalsanzahl bei gleichzeitiger Beschränkung der Irrtumsrate gering hält.

KAPITEL II

DIE REJECT ENTSCHEIDUNG

Bis hierhin haben wir stets die Einordnung aller beobachteten Muster in eine der vorgegebenen Klassen verlangt. Dieser Zwang zur Klassifizierung ist aber in medizinischen Diagnose - verfahren häufig nicht erwünscht. Stattdessen ist man bemüht, die schwierig zu klassifizierenden Fälle lieber vorerst nicht zuzuordnen, um sie z.B. einer zusätzlichen Kontrolle zu unterwerfen. Dazu wurden verschiedene Verfahren konstruiert (siehe [41], [43]), in der letzten Zeit vor allem das der "fuzzy diagnosis" (SMETS et al., [40]). Allen gemeinsam ist jedoch, daß die Auswirkungen der Verfahren auf die Fehler - kennungsrisiken nicht absehbar waren.

Das ist anders beim "Reject"-Verfahren. Diese Klassifizierungs - regel, die bis jetzt in der Medizin noch nicht angewandt worden ist, sieht für kritisch einzuordnende Fälle eine o-te Klasse "nicht erkannt" oder Reject-Klasse vor und ermöglicht damit, das Risiko der Fehlerkennung von vorneherein unter Kontrolle zu halten.

1. Reject- und Fehlklassifizierungsrisiko

Sei $\Omega = \{0, \dots, N\}$, $I = \{1, \dots, N\} = \Omega$ und $L: I \times \Omega \rightarrow \mathbb{R}$. Für $j \in I$ bezeichne $K_j = L(j, 0)$ den Verlust, den man beim Zurückweisen eines Musters der j -ten Klasse erleidet. Weiter sei für

jede randomisierte Entscheidungsfunktion $\varphi: \Omega \times \mathcal{X} \rightarrow [0, 1]$

$$e_x(\varphi) = \sum_{j \in I} L(j, i) \varphi(i|x) \gamma_x(j)$$
 das a posteriori Risiko der

Fehlklassifizierung von φ , gegeben $x \in \mathcal{X}$ und

$$r_x(\varphi) = \varphi(0|x) K(x) \quad \text{mit} \quad K(x) = \sum_{j \in I} K_j \gamma_x(j)$$
 das a posteriori

Risiko des Rejects von φ , gegeben $x \in \mathcal{X}$. Dann wird

$$r_x(\varphi) = e_x(\varphi) + r_x(\varphi) \quad \text{und} \quad \mathcal{R}(\varphi) = E(\varphi) + R(\varphi) \quad \text{mit}$$

$$E(\varphi) = \int_{\mathcal{X}} e_x(\varphi) f(x) dx \quad (\text{Fehlklassifizierungsrisiko}) \quad \text{und}$$

$$R(\varphi) = \int_{\mathcal{X}} r_x(\varphi) f(x) dx \quad (\text{Rejectrisiko}).$$

1.1 SATZ

Für jede nicht randomisierte Bayes-Entscheidung $\delta_K: \mathcal{O} \times \mathcal{X} \rightarrow \{0,1\}$

$$\text{gilt } E = E(\delta_K) \leq \bar{K} = E_Y(K) = \sum_{j \in I} p_j K_j.$$

δ_K heißt Reject-Entscheidung.

Beweis: Setzt man für $i \in \mathcal{O}$ und $\varphi^{(i)}(i|x) = 1 \quad \forall x \in \mathcal{X}$

$$r_x(i) = r_x(\varphi^{(i)}) = \sum_{j \in I} L(j,i) Y_x(j), \text{ so gilt für } x \in \Omega_i$$

$$r_x(i) \leq r_x(a) \quad \forall a \in \mathcal{O} \quad \text{und}$$

$$e(x) = e_x(\delta_K) = \begin{cases} r_x(i), & \text{falls } x \in \Omega_i \\ 0, & \text{falls } x \in \Omega_0 \end{cases} \quad (i \in I). \quad \text{Daraus folgt}$$

$$E(\delta_K) = \int_{\mathcal{X}} e(x) f(x) dx = \sum_{i \in I} \int_{\Omega_i} r_x(i) f(x) dx$$

$$\leq \int_{\mathcal{X}} r_x(0) f(x) dx = \int_{\mathcal{X}} \sum_{j \in I} p_j K_j f_j(x) dx = \sum_{j \in I} p_j K_j. \quad \cdot$$

Außerdem bekommt man, da δ_K Bayes-Lösung ist und da

$$R(\varphi) + E(\varphi) = \mathcal{S}(\varphi) \gg \mathcal{S}(\delta_K) = R(\delta_K) + E(\delta_K) \text{ ist,}$$

1.2 SATZ

Für jede randomisierte Entscheidungsfunktion $\varphi: \mathcal{O} \times \mathcal{X} \rightarrow [0,1]$

folgt aus $R(\varphi) \leq R(\delta_K) \quad E(\varphi) \gg E(\delta_K)$ und aus $E(\varphi) \leq E(\delta_K)$

$$R(\varphi) \gg R(\delta_K).$$

Somit gibt es keine Entscheidungsfunktion, die gleichzeitig ein besseres Fehlerkennungs- und Rejectrisiko besitzt als die Reject-Entscheidung. Hierfür definieren wir jetzt

die Rejectmenge $V_R(L) = \{x: \delta_K(o|x) = 1\} \subseteq \{x: \tau_x(o) \leq \tau(x)\}$

und die Akzeptiermenge $V_A(L) = \mathcal{X} \setminus V_R(L) \subseteq \{x: \tau_x(o) \geq \tau(x)\}$

mit $\tau(x) = \min_{j \in I} \tau_x(j)$. Damit ist

$$E(L) = E(\delta_K; L) = \int_{V_A(L)} \tau(x) f(x) dx \quad \text{und} \quad R(L) = R(\delta_K; L) = \int_{V_R(L)} K(x) f(x) dx.$$

Vergleicht man bei gleicher Verlustfunktion L auf $I \times I$ die Risiken einer Reject-Entscheidung mit dem einer gewöhnlichen Bayes-Entscheidung $\delta^*(O_1 = I)$, erhält man

1.3 SATZ

$$(1) \quad g(\delta^*) - E(\delta_K) = \int_{V_R(L)} f(x) \tau(x) dx \geq 0$$

$$(2) \quad g(\delta^*) - g(\delta_K) = \int_{V_R(L)} (\tau(x) - K(x)) f(x) dx \geq 0$$

Beweis:

$$\begin{aligned} (1) \quad g(\delta^*) - E(\delta_K) &= \int_{\mathcal{X}} \tau(x) f(x) dx - \int_{V_A(L)} \tau(x) f(x) dx \\ &= \int_{V_R(L)} \tau(x) f(x) dx. \end{aligned}$$

$$(2) \quad g(\delta^*) - g(\delta_K) = g(\delta^*) - E(\delta_K) - R(\delta_K)$$

$$= \int_{V_R(L)} f(x) (\tau(x) - K(x)) dx \geq 0, \text{ weil für } x \in V_R(L)$$

$K(x) \leq \tau(x)$ gilt. $\blacksquare$

Während die Rejectregel also geringere Irrtumsrisiken ergibt, liefert sie gleichzeitig, wie unten gezeigt wird, auch eine kleinere Rate der richtigen Zuordnung. Man kann jetzt auch, analog zu CHU[6] (vgl. auch CHOW[4]) bei beschränktem Fehlerkennungsrisiko die Entscheidungsfunktionen mit minimaler Rejectrate suchen. Von diesen Entscheidungsfunktionen "of the minimum rejection type" (m.r.t.) ist diejenige "best possible" (b.p.), die das kleinste $E(\varphi)$ besitzt.

Zu $\alpha = E(\delta_K)$ ist dann δ_K selbst unter allen φ mit $E(\varphi) \leq \alpha$ b.p., da aus $E(\varphi) \leq \alpha$ $R(\delta_K) \leq R(\varphi)$, also δ_K m.r.t., und da aus $E(\varphi) < E(\delta_K)$ $R(\varphi) > \alpha$, also δ_K b.p. folgt (nach 1.2). Für $L(j,i) = 1 - \delta_{ij}$ auf $I \times I$ und $K_j = c$ für alle $j \in I$ zeigte CHU [6], daß auch zu beliebigem $\alpha > 0$ eine b.p. Entscheidungs-funktion existiert und daß diese i.a. randomisiert ist. Umgekehrt kann man auch "decision functions of minimum error type" behandeln. Diese sind jedoch für den medizinischen Klassifizierungsprozeß weniger geeignet, da hier die Beschränkung der Zahl der Fehldiagnosen im Vordergrund steht. Da die Reject-Regel ein mehrfaches Gewichten jeder einzelnen Klasse (durch $L(j,i)$) erlaubt und das totale Risiko unter Kontrolle hält, stellt sie auch eine brauchbare Alternative zu dem Verfahren von Neyman-Pearson dar.

2. Spezielle Verlustfunktionen

Ist wieder $L(j,i) = L_j(1 - \delta_{ij})$ für alle $i, j \in I$, so gilt mit $L(x) = \sum_{j \in I} L_j \gamma_x(j)$ und $m(x) = \max_{j \in I} (\gamma_x(j) L_j)$ $\{x: m(x) > L(x) - K(x)\} \subseteq V_A(L) \subseteq \{x: m(x) \geq L(x) - K(x)\}$ und $\{x: m(x) < L(x) - K(x)\} \subseteq V_R(L) \subseteq \{x: m(x) \leq L(x) - K(x)\}$. Damit bekommt man

2.1 SATZ

Sei δ_K eine Reject-Entscheidung zu γ und $L = (L_1, \dots, L_N)$, $K = (K_1, \dots, K_N)$.

Dann gibt es eine a priori Verteilung γ' , ein $K' = (K_1', \dots, K_N')$ und eine zugehörige Reject-Entscheidung $\delta_{K'}$, so daß gilt

- (i) $L_j' = 1$ für alle $j \in I$
- (ii) $V_A(\delta_{K'}) = V_A(\delta_K)$ und $V_R(\delta_{K'}) = V_R(\delta_K)$
- (iii) $E(\delta_K) = \bar{L} E(\delta_{K'})$ und $R(\delta_K) = \bar{L} R(\delta_{K'})$

Dabei ist $p_j' = p_j L_j / L$ und $K_j' = K_j / L_j$.

Beweis: Seien γ' und K' wie angegeben definiert, außerdem

$$f'(x) = \sum_{j \in I} p_j' f_j(x) \text{ und } L_j' = 1. \text{ Dann folgt}$$

$$f(x)L(x) = \bar{L} L'(x)f'(x) \text{ und } f(x)K(x) = \sum_{j \in I} p_j K_j f_j(x)$$

$$= \bar{L} \sum_{j \in I} (p_j L_j / \bar{L}) (K_j / L_j) f_j(x) = \bar{L} f'(x) K'(x).$$

Sei δ_K' eine Entscheidungsfunktion mit

$$\delta_K'(o|x) = 1 \iff \delta_K(o|x) = 1 \quad \text{und}$$

$$\delta_K'(j|x) = 1 \Rightarrow m'(x)f'(x) = p_j' L_j' f_j(x).$$

Dann ist sicher $V_R(\delta_K') = V_R(\delta_K)$. Außerdem ist δ_K' eine

Reject-Regel, weil für $x \in V_R(\delta_K)$ $K(x) \leq L(x) - m(x)$, also

$$\bar{L} f'(x) K'(x) \leq \bar{L} f'(x) (L'(x) - m'(x)) \text{ und analog für } x \in V_A(\delta_K)$$

$K'(x) \geq L'(x) - m'(x)$ gilt. Für die Risiken folgt dann

$$E(\delta_K') = \int_{V_A(\delta_K)} f(x)(L(x) - m(x)) dx = \bar{L} \int_{V_A(\delta_K')} (L'(x) - m'(x)) f'(x) dx$$

$$\text{und } R(\delta_K') = \int_{V_R(\delta_K')} K(x)f(x) dx = \bar{L} \int_{V_R(\delta_K')} K'(x)f'(x) dx.$$

2.2 BEMERKUNGEN

(1) Nach 2.1 kann man nun $L(j,i) = 1 - \delta_{ij}$ auf $I \times I$ annehmen.

$$\text{Setzt man dafür } C(\delta_K) = \int_{V_A(\delta_K)} m(x)f(x) dx = \int_{V_A(\delta_K)} f(x) - E(\delta_K),$$

so stellt $C(\delta_K)$ den erwarteten Wert der richtigen Zuordnung

dar. Für die gewöhnliche Bayes-Entscheidung δ^* entspricht dem

$$C(\delta^*) = \int_{\mathcal{X}} m(x)f(x) dx = 1 - S(\delta^*). \text{ Da } C(\delta^*) - C(\delta_K) = \int_{V_R(\delta_K)} m(x)f(x) dx$$

ist, ergibt eine Reject-Entscheidung neben einer geringeren

Irrtumsrate (nach 1.3) auch eine niedrigere Rate von richtigen

Zuordnungen. In einen Reject werden also sowohl möglicher -

weise richtige als auch falsche Zuordnungen umgewandelt. Dabei wird der Verlust in C meist größer sein als der in S, da

$$(C(\delta^*) - C(\delta_K)) - (S(\delta^*) - S(\delta_K)) = \int_{V_K(\delta_K)} (2m(x) + K(x) - 1) f(x) dx.$$

Dies ist sicher ≥ 0 für $N=2$, da dann $m(x) \geq 1/2 \quad \forall x \in X$ gilt.

$$(2) \text{ Für } N=2 \text{ gilt für } x \in V_R(L) \quad \frac{K}{L_1 - K_1} \leq \frac{p_1 f_1(x)}{p_2 f_2(x)} \leq \frac{L_2 - K}{L_2 - K_1}.$$

Also ist $V_R(L) = \emptyset$, falls $\frac{K}{L_1 - K_1} > \frac{L_2 - K}{L_2 - K_1}$ oder

falls $L_1(\alpha L_2 - K_2) + L_2((1-\alpha)L_1 - K_1) \leq 0 \quad (0 \leq \alpha \leq 1)$ erfüllt ist.

2.3 PAARWEISE DARSTELLUNG DER REJECTWAHRSCHEINLICHKEIT

Dazu zerlegen wir (bei $L(j,i) = 1 - \delta_{ij}$) den Merkmalsraum X in paarweise disjunkte Mengen S_{ij} ($i < j$) mit

$S_{ij} \subseteq \{x: p_i f_i(x), p_j f_j(x) > p_k f_k(x) \quad \forall k \neq i, j\}$. Dann er-

hält man für die Rejectwahrscheinlichkeit $P(V_R(K)) = \int_{V_R(K)} f(x) dx$:

$$P(V_R(K)) = \sum_{i < j, i, j \in I} (p_i + p_j) P_{ij}(V_R^{(i,j)}(K_i, K_j)) + \varepsilon(K).$$

Dabei ist $P_{ij}(V_R^{(i,j)}(K_i, K_j)) = \int_{V_R^{(i,j)}(K_i, K_j)} \left(\frac{p_i f_i(x)}{p_i + p_j} + \frac{p_j f_j(x)}{p_i + p_j} \right) dx$

die Rejectwahrscheinlichkeit zum 2-Klassen-Fall (i, j) .

Für $\varepsilon(K) = \sum_{i < j} \varepsilon_{ij}(K)$ ergibt sich $\varepsilon_{ij}(K) = \varepsilon_{ij}^{(1)} + \varepsilon_{ij}^{(2)} + \varepsilon_{ij}^{(3)}$ mit (1)

$$\varepsilon_{ij}^{(1)} = \int_{S_{ij} \cap V_R(K)} (f(x) - (p_i f_i(x) + p_j f_j(x))) dx \text{ wird dabei durch}$$

die Werte ab der drittgrößten Dichte bestimmt und kann somit durch die BK-3, BK-4 u.s.w. abgeschätzt werden.

$$(2) \quad \varepsilon_{ij}^{(2)} = -P_{ij}((V_R^{(i,j)}(K) \cap S_{ij}) \setminus (V_R(K) \cap S_{ij})) \cdot (p_i + p_j)$$

ist dagegen schwieriger abzuschätzen. Jedoch ist zu erkennen, daß auch $\varepsilon_{ij}^{(2)}$ von der Größe der f_k ($k \neq i, j$) auf S_{ij} abhängt (vgl. I, 6.7(1)).

$$(3) \varepsilon_{ij}^{(3)} = - \int_{\bar{S}_{ij} \cap V_R^{(i,j)}(K)} (p_i f_i(x) + p_j f_j(x)) dx$$

Für $x \in \bar{S}_{ij} \cap V_R^{(i,j)}(K)$ folgt nach 2.2(2) aus $p_i f_i(x) > a$ $p_j f_j(x) > K_i a / (1 - K_j)$. Also können auf $\bar{S}_{ij} \cap V_R^{(i,j)}(K)$ nur $p_i f_i(x)$ und $p_j f_j(x)$ entweder gemeinsam einen Beitrag liefern oder gemeinsam klein sein. Die zweite Möglichkeit ist dann gegeben, wenn man generell verlangt, daß jeweils nur zwei Dichten wesentlich zu $f(x)$ beitragen. In diesem Fall ist, wie in (1) und (2), auch $\varepsilon_{ij}^{(3)}$ zu vernachlässigen.

Stellt man die Rejectwahrscheinlichkeit (analog I,1.2) durch $P(V_R(K)) = \sum_{i < j} P(V_R^{(i,j)}(K)) + \varepsilon_0(K)$ dar, reduzieren sich die $\varepsilon_{ij;0}^{(1)}$ auf Ausdrücke des Typs (2).

3. Error-Reject-Tradeoff

In diesem Paragraphen soll die Verlustfunktion auf $I \times I$ wieder beliebig, der Verlust durch einen Reject aber für alle Klassen gleich gewichtet sein (also $K_j = t$ für alle $j \in I$). Nach 1.1 stellt dann t eine obere Schranke für das Irrtums-Risiko dar. Wir beschränken uns jetzt auf Reject-Entscheidungen, die im Zweifel einer Zurückweisung den Vorrang geben, also $\delta_t(0|x) = 1 \Leftrightarrow \gamma(x) > t$. Die Reject- bzw. Akzeptiermengen sind dann monoton fallende bzw. wachsende Funktionen von t . Außerdem ist die Akzeptierwahrscheinlichkeit $P(t) = P(V_A(t)) = \int_{\{x: \gamma(x) < t\}} f(x) dx$ eine monoton wachsende, linksstetige Verteilungsfunktion.

3.1 SATZ

Sei $L: I \times \Omega \rightarrow \mathbb{R}$ mit $K_j = L(j, \omega) = t$ für alle $j \in I$. Dann läßt sich das Fehlerkennungsrisiko darstellen als

$$E(t) = E(L) = \int_0^t t' dP(t') .$$

Beweis: Sei $K': \Delta K_j = K_j - K_j' > 0$. Dann folgt aus $K(x) > K'(x)$ für alle $x \in \mathcal{X}$ $V_R(K) \subseteq V_R(K')$ und $V_A(K) \supseteq V_A(K')$. Sei $dV_A = V_A(K) \setminus V_A(K') \subseteq \{x: K(x) > \tau(x) > K'(x)\}$ und $dE := E(K) - E(K') = \int_{dV_A} \tau(x) f(x) dx$. Damit gilt

$$\int_{dV_A} f(x) K'(x) dx \leq dE \leq \int_{dV_A} f(x) K(x) dx , \text{ also}$$

$$\sum_{j \in I} p_j K_j P_j(dV_A) - \sum_{j \in I} p_j \Delta K_j P_j(dV_A) \leq dE \leq \sum_{j \in I} p_j K_j P_j(dV_A)$$

(mit $P_j(A) = \int_A f_j(x) dx$). Sei $K_j = t$ und $\Delta K_j = \Delta t = t/m$ (für $n \geq m > 0$ und $j \in I$), außerdem $a_1 = \frac{1}{m} \cdot t$ für $0 \leq l \leq m$. Dann gilt

$$t P(dV_A) \geq dE \geq t P(dV_A) - \Delta t P(dV_A).$$

$$\text{Da } E(t) = E(t) - E(0) = \sum_{l=1}^m (E(a_l) - E(a_{l-1})) \text{ und}$$

$$a_1 dP(a_1) - \Delta t dP(a_1) \leq E(a_1) - E(a_{1-1}) \leq a_1 dP(a_1)$$

mit $dP(a_1) = P(V_A(a_1) \setminus V_A(a_{1-1}))$ gilt, folgt

$$\sum_{l=1}^m a_1 dP(a_1) - \sum_{l=1}^m \Delta t dP(a_1) \leq E(t) \leq \sum_{l=1}^m a_1 dP(a_1).$$

Weil $\sum_{l=1}^m \Delta t dP(a_1) = \Delta t (P(V_A(t)) - P(V_A(0)))$ gegen 0 und

$$\sum_{l=1}^m a_1 dP(a_1) \text{ gegen } \int_0^t t' dP(t') \text{ für } m \text{ gegen } \infty \text{ konvergiert, folgt die Behauptung.}$$

Damit bekommt man

$$E(t) = - \int_0^t t' dP(V_R(t')) \quad (\text{Error - Reject - Tradeoff})$$

und so eine Verallgemeinerung der Ergebnisse von CHOW [5]

auf den Fall beliebiger Verlustfunktion auf $I \times I$.

Ist $P(t)$ λ -stetig (λ Borel-Maß auf $\mathbb{R}$), wie etwa bei normal - verteilten Dichten, kann man nach dem Satz von Radon-Nikodym

schreiben: $E(t) = \int_0^t t' g(t') d\lambda(t')$. Dabei ist g die zu -

gehörige Dichte (bzgl. λ). Ist P sogar differentierbar nach t ,

kann man partiell integrieren und bekommt

$$E(t) = \int_0^t t' \frac{dP(t')}{dt'} dt' = t P(t) - \int_0^t P(t') dt'.$$

Für $t=1$ und $L=1-\delta_{ij}$ auf $I \times I$ ist dann

$$g = E(1) = 1 - \int_0^1 P(t) dt = \int_0^1 P(V_R(t)) dt.$$

In diesem Fall erhält man beim Vergleich zweier Klassenein - teilungen Γ_1 und Γ_2

3.2 SATZ

Für $\Gamma = \Gamma_1, \Gamma_2$ mit $L^\Gamma(j,i)=1-\delta_{ij}$ auf $I(\Gamma) \times I(\Gamma)$ seien $P^\Gamma(t)$ differentierbar auf $(0,1)$.

Ist dann Γ_1 t -besser als Γ_2 (d.h. $\frac{P(V_{R_1}^\Gamma(t))}{J(\Gamma_1)} \leq \frac{P(V_{R_2}^\Gamma(t))}{J(\Gamma_2)}$)

für alle $0 \leq t \leq 1$, so ist Γ_1 auch Q -besser als Γ_2 .

3.3 SATZ

Für die Klasseneinteilung $\Gamma = \{J, \{k\}; k \notin J\}$ gilt (bei $L(j,i)=1-\delta_{ij}$)

$$R^{\Gamma_0}(t) - R^\Gamma(t) = t(P(V_{R_0}^\Gamma(t)) - P(V_R^\Gamma(t)))$$

$$= t \sum_{i \in J, j \notin J} (p_i + p_j) P_{ij}(V_R^{(i,j)}(t)) + t \varepsilon^\Gamma(t).$$

Beweis: analog zu I, Lemma 1.1 unter Benutzung der paarweisen Darstellung der Rejectwahrscheinlichkeit in 2.3. ■

Auch hier ist $\varepsilon^\Gamma(t) < 1$, falls für alle $x \in \mathbb{R}$ höchstens zwei Dichten jeweils wesentlich zu $f(x)$ beitragen.

3.4 BEISPIELE

Sei $N=2$ und $p_1=p_2=1/2$

(1) $f_j \sim \mathcal{N}(\mu_j, \sigma)$ für $j=1,2$ und $\mu_1 > \mu_2$. Dann folgt aus $m(x) > 1/2$ für alle $x \in \mathbb{R}$ $P(V_R(t)) = 0$ für alle $t > 1/2$. Für $0 < t \leq 1/2$

bekommt man (siehe [5]) $P(V_R(t)) = \phi(b) - \phi(a)$ mit

$$\phi(z) = (2\pi)^{-1/2} \int_{-\infty}^z e^{-x^2/2} dx,$$

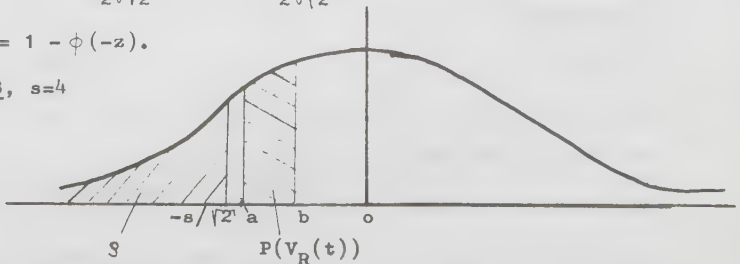
$a = -s/2 - (\ln(1/t-1))/s$, $b = -s/2 + (\ln(1/t-1))/s$ und $s = (\mu_1 - \mu_2)/\sigma$.

Dagegen gilt für das Bayes-Risiko (nach [45])

$$s = \frac{1}{2} \left(1 - \phi\left(\frac{\mu_1 - \mu_2}{2\sigma\sqrt{2}}\right) + \phi\left(\frac{\mu_2 - \mu_1}{2\sigma\sqrt{2}}\right) - 1 \right) = 1 - \phi(s/\sqrt{2}),$$

da $\phi(z) = 1 - \phi(-z)$.

$$t=1/8, s=4$$



Vergleicht man zwei solche Paare (f_1^1, f_2^1) und (f_1^2, f_2^2) mit den zugehörigen $s^j = (\mu_1^j - \mu_2^j)/\sigma^j$, sieht man, daß

$$s^1 > s^2 \Leftrightarrow s^1 \leq s^2 \Leftrightarrow P(V_R^1(t)) > P(V_R^2(t))$$

für ein (oder alle) $t \in (0, 1/2)$.

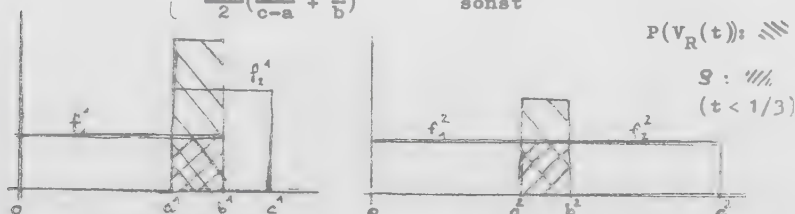
Ebenso erhält man beim Vergleich zweier Γ_1 und Γ_2

$$\Gamma_1 \stackrel{Q}{\prec} \Gamma_2 \Leftrightarrow \Gamma_1 \stackrel{t}{\prec} \Gamma_2 \quad \text{für ein } t \in (0, 1/2),$$

damit gilt für diesen einfachsten Typ von Dichten auch die Umkehrung von Satz 3.2.

(2) Daß dies nicht immer der Fall ist, zeigt folgendes Beispiel: Seien g_1 bzw. g_2 auf $(0, b]$ bzw. $(a, c]$ gleichverteilte Dichten. Dann ist bei $b > a$ $\vartheta = (b-a)/2b$ und

$$P(V_R(t)) = \begin{cases} 0 & , \text{ falls } \frac{b}{c-a} \geq \frac{1-t}{t} \\ \frac{b-a}{2} \left(\frac{1}{c-a} + \frac{1}{b} \right) & \text{sonst} \end{cases}$$



Aus den Diagrammen erkennt man, daß in beiden Fällen die Bayes-Risiken gleich sind, die Rejectwahrscheinlichkeiten aber für einige $t \in [0, 1]$ stark voneinander abweichen. Auch Satz 3.2 gilt jetzt nicht.

Die so gezeigte Tatsache, daß die Beurteilung der Güte einer Klasseneinteilung wesentlich von der Wahl der Entscheidungs - funktion abhängt, wird im nächsten Kapitel noch am Beispiel einer nicht parametrischen Entscheidungsregel demonstriert.

KAPITEL III

DIE NEAREST - NEIGHBOR - ENTSCHEIDUNG

Bei dieser nicht parametrischen Entscheidungsregel sei wieder $\Omega = I = \Omega$ und $L: I \times I \rightarrow \mathbb{R}$ vorausgesetzt.

Außerdem sei eine Metrik $d: X \times X \rightarrow \mathbb{R}^+$ gegeben.

1 DEFINITION

Für $x^1, \dots, x^n \in X \setminus \{0\}$ mit $n \in \mathbb{N}$ seien die richtigen Zuordnungen $\psi_n(x^1), \dots, \psi_n(x^n) \in I$ bekannt. Die Nearest-Neighbor-Rule (NNR)

$\psi_n: X \rightarrow I$ ordnet jedem neu beobachteten Wert $x \in X$ die Klasse zu, der der nächste der $x^1, \dots, x^n$ angehört:

$\psi_n(x) = \psi_n(\tilde{x}^n)$ für $\tilde{x}^n \in M(x) = \{y \in \{x^1, \dots, x^n\} : d(x, y) = \min_{j=1, \dots, n} d(x, x^j)\}$, falls $\psi_n(\tilde{x}^n) \leq \psi_n(y)$ für alle $y \in M$.

$\tilde{x}^n$ heißt Nearest-Neighbor von x ($NN(n)$).

Sind p_j und f_j ($j \in I$) die jetzt unbekannten zugrunde liegenden a priori Wahrscheinlichkeiten und bedingten Dichten, so sind $x^1, \dots, x^n, x$ identisch nach $f(x)$ verteilt. Hierfür bewiesen COVER und HART [10]:

2 LEMMA

Ist X separabel bezüglich der Metrik d , so konvergieren für n gegen ∞ die $NN(n)$ $\tilde{x}^n$ P-fast sicher gegen x .

Also ist $P \left\{ \lim_{n \rightarrow \infty} \tilde{x}^n = x \right\} = 1$. Außerdem zeigten COVER und HART [10], daß das a posteriori Risiko von ψ_n , gegeben x ,

$$\tau_x(\psi_n) = \sum_{i,j \in I} L(j,i) \gamma_x(j) \gamma_{\tilde{x}^n}(i) \quad \text{ist.}$$

Setzt man $\psi_n(i|x) = \gamma_{\tilde{x}^n}(i)$, ist ψ_n eine randomisierte Entscheidungsregel mit $\tau^n(x) := \tau_x(\psi_n) = \tau_x(\psi_n)$ und

$g^n = s(\psi_n) = s(\psi_n)$, dem durchschnittlichen Risiko der $NN(n)$.

Von besonderem Interesse ist das asymptotische Verhalten von $r^n(x)$ und g^n . Für $\vec{g} = \lim_{n \rightarrow \infty} g^n$ erhielten COVER und HART [10] einen bemerkenswerten Satz:

3 SATZ

Sei X ein separabler Raum bezüglich einer Metrik d mit einem Wahrscheinlichkeitsmaß ν . $f_1, \dots, f_N$ seien Dichten bzgl. ν , so daß für fast alle $x \in X$ gilt

- (i) x ist ein Stetigkeitspunkt von $f_1, \dots, f_N$ oder
- (ii) $\nu\{x\} \neq 0$.

Dann existiert das NN-Irrtumsrisiko $\vec{g} = \lim_{n \rightarrow \infty} g^n$ (für $L(j, i) = 1 - \delta_{ij}$) und hat folgende Schranken:

$$g \leq \vec{g} \leq g(2 - \frac{N}{N-1} g).$$

In unseren Anwendungsbeispielen werden nur Dichten bzgl. des Borelmaßes λ^S auf $\mathbb{R}^S$ betrachtet. Hierfür ist die Bedingung (ii) sicher nicht erfüllt. Andererseits sind die in Frage kommenden Dichten (Normalverteilungen und ihre Mischungen) stetige Funktionen bzgl. der durch die euklidische Metrik erzeugten Topologie. Erzeugt die Metrik d eine feinere Topologie als diese, sind die $f_1, \dots, f_N$ auch stetige Funktionen von $(X, d) \rightarrow \mathbb{R}$. Hierfür gilt dann auch, da $\tilde{x}^n \rightarrow x$ f.u.,

$$\vec{r}(x) = \lim_{n \rightarrow \infty} r^n(x) = \lim_{n \rightarrow \infty} \left(\sum_{\substack{j=1 \\ j \neq i}}^N \gamma_x(j) \gamma_{\tilde{x}^n}(i) \right) = \sum_{\substack{j=1 \\ j \neq i}}^N \gamma_x(j) \gamma_x(i) \quad \text{f.u.}$$

und nach Anwendung des Satzes von der majorisierten Konvergenz

$$\vec{g} = \int_X \vec{r}(x) f(x) dx = \sum_{\substack{j=1 \\ j \neq i}}^N \int_X \frac{p_j p_{j-1}(x) f_j(x)}{f(x)} dx.$$

Damit bekommt man eine Beziehung zu den Bhattacharyya-Koeffizienten:

4 SATZ

Für $N = 2$ gilt

$$(1) E(\sqrt{\tilde{\tau}(x)}) = \sqrt{2p_1p_2} \beta$$

$$(2) \tilde{g} \geq 2p_1p_2\beta^2$$

Beweis: (1) folgt direkt aus

$$f(x) \sqrt{\tilde{\tau}(x)} = \sqrt{2p_1p_2f_1(x)f_2(x)},$$

(2) mit der Hölder'schen Ungleichung

$$\begin{aligned} \beta &= \int_{\mathcal{X}} \sqrt{f_1(x)f_2(x)} dx = \int_{\mathcal{X}} \sqrt{\frac{f_1(x)f_2(x)}{f(x)}} \sqrt{f(x)} dx \\ &\leq \sqrt{\int_{\mathcal{X}} \frac{f_1(x)f_2(x)}{f(x)} dx} \sqrt{\int_{\mathcal{X}} f(x) dx} = \sqrt{\frac{\tilde{g}}{2p_1p_2}}. \quad \blacksquare \end{aligned}$$

5 FOLGERUNG

Unter den Voraussetzungen von Satz 3 gilt für beliebiges N

$$g \geq \frac{N-1}{N} \left(1 - \sqrt{1 - \frac{2N}{N-1} \left(\sum_{\substack{i < j \\ i \in I}} p_i p_j \beta_{ij}^2 \right)} \right) \geq \sum_{\substack{i < j \\ i \in I}} p_i p_j \beta_{ij}^2.$$

Beweis: Nach 4(2) erhält man

$$\tilde{g} = \sum_{i \neq j} p_i p_j \int_{\mathcal{X}} \frac{f_i(x)f_j(x)}{f(x)} dx \geq \sum_{i \neq j} p_i p_j \beta_{ij}^2$$

$$\text{und, da nach S.3 } g \geq \frac{N-1}{N} \left(1 - \sqrt{1 - \frac{N}{N-1} \tilde{g}} \right) \text{ ist,}$$

die Behauptung. $\blacksquare$

Für $N=2$ ist das die Aussage von I,2.3. Man kann jetzt auch angeben, wodurch der Fehler in der Abschätzung in Folg.5 bedingt ist:

6 SATZ

Unter den Voraussetzungen von Satz 3 gilt mit $\beta(I) = \sum_{i < j} p_i p_j \beta_{ij}^2$

$$g = \frac{N-1}{N} \left(1 - \sqrt{1 - \frac{2N}{N-1} \beta(I) - \frac{N}{N-1} \left(2 \sum_{\substack{i < j \\ i \in I}} V(\tilde{\tau}_{ij}(x)) + \frac{N}{N-1} V(\tau(x)) \right)} \right)$$

Dabei ist $\tilde{c}_{ij}(x) = 2p_i p_j \frac{f_1(x)f_j(x)}{f^2(x)}$ und $V(\cdot) = \text{Var}(\cdot)$.

Beweis: Setzt man noch

$\tilde{s}_{ij} = E(\tilde{c}_{ij}(x))$, so hat man $\tilde{c}(x) = \sum_{i < j} \tilde{c}_{ij}(x)$ und $\tilde{s} = \sum_{i < j} \tilde{s}_{ij}$,
außerdem $0 \leq V(\sqrt{\tilde{c}_{ij}(x)}) = E(\tilde{c}_{ij}(x)) - (E(\sqrt{\tilde{c}_{ij}(x)}))^2 = \tilde{s}_{ij} - 2p_i p_j f_{ij}^2$

also $\tilde{s} = \sum_{i < j} V(\sqrt{\tilde{c}_{ij}(x)}) + 2 \beta(I)$. Nun gilt nach [10]

$$\tilde{s} = 2s - \frac{N}{N-1} V(\tilde{c}(x)) - \frac{N}{N-1} s^2 \quad \text{und daher}$$

$$s = \frac{N-1}{N} \left(1 - \sqrt{1 - \frac{N}{N-1} \tilde{s} - \left(\frac{N}{N-1}\right)^2 V(\tilde{c}(x))} \right)$$

Durch Einsetzen von $\tilde{s}$ folgt die Behauptung. ■

7 BEMERKUNGEN

(1) Wählt man die Metrik d so, daß die $V(\sqrt{\tilde{c}_{ij}(x)})$ klein werden, wird s also bei kleinen Schwankungen im a posteriori Bayes-Risiko im wesentlichen durch die untere Schranke $d_1(s)$ bestimmt. Allerdings ist schon an numerischen Beispielen in I,7.1 gezeigt worden, daß ein kleines $V(\tilde{c}(x))$ selten erwartet werden kann.

(2) Die Gleichung $\tilde{s} = \sum_{i < j} \tilde{s}_{ij}$ aus dem Beweis von S.6 ergibt nun wieder eine paarweise Darstellung des NN-Irrtums. Dabei sind, ähnlich I,1.2 (vgl. auch II,2.3) die $\tilde{s}_{ij}$ selbst keine NN-Risiken, sondern es gilt

$$\tilde{s}(i,j) = \int_{\mathcal{X}} 2p_i p_j \frac{f_1(x)f_j(x)}{p_i f_1(x) + p_j f_j(x)} dx \geq \tilde{s}_{ij}.$$

Für den Übergang von Γ_0 zu einer gröberen Zerlegung $\Gamma = \{I_1, \dots, I_M\}$ bekommt man also

$$\tilde{s} - \tilde{s}(\Gamma) = \sum_{l=1}^M \sum_{\substack{k,j \in I_l \\ k < j}} \tilde{s}(k,j) - \varepsilon_{NN}(\Gamma)$$

$$\text{mit } \varepsilon_{NN}(\tau) = \sum_{i=1}^M \sum_{\substack{k+j \\ k \in I_1}} (\vec{s}(k,j) - \vec{s}_{kj}) \geq 0 \text{ (wie in I,1.6)}$$

(3) Für die praktische Anwendung der NN-Regel spielt natürlich die Auswahl eines geeigneten Abstandsmaßes die größte Rolle. Dabei wiederum ist das Konvergenzverhalten von $\tilde{x}^n$, $\tau^n(x)$ und g^n von großer Bedeutung. Dieses ist u.a. von COVER[9] und PETERSON [35] untersucht worden.

Die k-NN-Regel, ein Verfahren, das die k-nächsten Nachbarn von x bei der Klassifizierung berücksichtigt, wurde von COVER und HART [10] in gleicher Weise untersucht wie die 1-NN-Regel. Darüberhinaus diente sie HELLMAN [18] dazu, ein NN-Verfahren mit Reject zu konstruieren. Es scheint jedoch so, als ob ein solches NN-Reject-Verfahren auch für die 1-NN-Regel zu konstruieren ist. Dabei könnten Kugeln vom Radius t um $x \in X$ die Akzeptiermenge bestimmen.

KAPITEL IV

MODIFIZIERTE ENTSCHEIDUNGSREGELN

1. Modifizierte Bayes-Entscheidung

In I,4.4 haben wir versucht, die Klasseneinteilung $\bar{\Gamma}$ zu finden, die $g(\bar{\Gamma}) + \gamma c(\bar{\Gamma})$ für alle Γ aus $\mathcal{K}_\alpha = \{\Gamma < \Gamma_0: \mathbb{I}(\Gamma) \geq \alpha\}$ (dort war $\gamma = s$ und $\alpha = 0$) minimiert. Dabei war $c(\Gamma) = \mathbb{I}^{-1}(\mathbb{I} - \mathbb{I}(\Gamma))$.

$$\text{Nun ist } g(\bar{\Gamma}) + \gamma c(\bar{\Gamma}) = \int_{\mathcal{X}} (\tau_x(\delta^{\bar{\Gamma}}) + \gamma c(\bar{\Gamma})) f(x) dx$$

$$= \min_{\Gamma \in \mathcal{K}_\alpha} \int_{\mathcal{X}} (\tau_x(\delta^{\Gamma}) + c(\Gamma)) f(x) dx$$

$$= \min_{\substack{\Psi: \mathcal{K}_\alpha \rightarrow [0,1] \\ \sum_{\Gamma \in \mathcal{K}_\alpha} \Psi(\Gamma) = 1}} \int_{\mathcal{X}} \sum_{\Gamma \in \mathcal{K}_\alpha} \Psi(\Gamma) (\tau_x(\delta^{\Gamma}) + \gamma c(\Gamma)) f(x) dx.$$

Dabei wird für die Verteilung Ψ auf $\mathcal{K}_\alpha$ mit Wahrscheinlichkeit $\Psi(\Gamma)$ die Klassenauswahl Γ getroffen. Machen wir diese Auswahl

der zu unterscheidenden Klassen von der Beobachtung $x \in \mathcal{X}$ abhängig, betrachten wir Elemente aus

$$\Psi(\mathcal{K}_\alpha) = \left\{ \Psi: \mathcal{K}_\alpha \times \mathcal{X} \rightarrow [0,1]: \sum_{\Gamma \in \mathcal{K}_\alpha} \Psi(\Gamma|x) = 1 \quad \forall x \in \mathcal{X} \right\}$$

$$\text{und für } \Psi \in \Psi(\mathcal{K}_\alpha) \text{ und } x \in \mathcal{X} \quad \sum_{\Gamma \in \mathcal{K}_\alpha} \Psi(\Gamma|x) (\tau_x(\delta^{\Gamma}) + \gamma c(\Gamma)).$$

Dies bedeutet, daß man für jedes $x \in \mathcal{X}$ mit Wahrscheinlichkeit $\Psi(\Gamma|x)$ die Zerlegung Γ auswählt und dann die Bayes-Entscheidung δ^{Γ} trifft. Dabei resultiert das a posteriori Risiko

$\tau(\Gamma|x) = \tau_x(\delta^{\Gamma})$. Die Kosten dieses Stoppens bei Γ stellt der gewichtete Verlust an Information $\gamma c(\Gamma)$ dar. Wir definieren

$$\mathcal{D}(\mathcal{K}_\alpha) = \left\{ \varphi = \{\varphi^{\Gamma}; \Gamma \in \mathcal{K}_\alpha, \varphi^{\Gamma} \in \mathcal{D}(\Gamma)\} \right\} \text{ und}$$

$r_x(\Psi, \varphi) = \sum_{\Gamma \in \mathcal{K}_\alpha} \Psi(\Gamma|x) (\tau_x(\varphi^{\Gamma}) + \gamma c(\Gamma))$, das modifizierte a posteriori Risiko von $(\Psi, \varphi) \in \Psi(\mathcal{K}_\alpha) \times \mathcal{D}(\mathcal{K}_\alpha)$, gegeben x .

Für jedes $\psi_0 \in \Psi(\mathcal{X}_d)$, das ein Γ mit minimalem $r(\Gamma|x) = \tau(\Gamma|x) + c(\Gamma)$ mit Wahrscheinlichkeit 1 auswählt, bekommt man

1.1 SATZ

Für jedes so definierte ψ_0 und für jedes

$\delta_0 \in \mathcal{D}^*(\mathcal{X}_d) = \{\varphi \in \mathcal{D}(\mathcal{X}_d) : \varphi^\Gamma \in \mathcal{D}^*(\Gamma)\}$ wird das a posteriori Risiko von (ψ, φ) , gegeben x , für alle $x \in \mathcal{X}$ in (ψ_0, δ_0) auf $\Psi(\mathcal{X}_d) \times \mathcal{D}(\mathcal{X}_d)$ minimal.

(ψ_0, δ_0) heißt modifizierte Bayes-Entscheidung (MBE).

Beweis:

$$\begin{aligned} r_x(\psi_0, \delta_0) &= \sum_{\Gamma \in \mathcal{X}_d} \psi_0(\Gamma|x) (\tau_x(\delta_0^\Gamma) + \gamma c(\Gamma)) \\ &= \min_{\Gamma \in \mathcal{X}_d} ((\tau_x(\delta_0^\Gamma) + \gamma c(\Gamma)) (\sum_{\Gamma \in \mathcal{X}_d} \psi(\Gamma|x))) \\ &\leq \sum_{\Gamma \in \mathcal{X}_d} \psi(\Gamma|x) (\tau_x(\varphi^\Gamma) + \gamma c(\Gamma)) \quad \text{für alle } (\psi, \varphi) \in \Psi(\mathcal{X}_d) \times \mathcal{D}(\mathcal{X}_d), \\ &\text{da } \delta_0^\Gamma \text{ Bayes-Entscheidung zu } I(\Gamma) \text{ ist.} \quad \blacksquare \end{aligned}$$

In diesem Beweis sieht man auch, daß

$r_x(\psi, \varphi) \geq r_x(\psi, \delta_0)$ für alle $\delta_0 \in \mathcal{D}^*(\mathcal{X}_d)$, $\varphi \in \mathcal{D}(\mathcal{X}_d)$ und $\psi \in \Psi(\mathcal{X}_d)$ gilt. Definiert man das durchschnittliche Risiko der modifizierten Entscheidungsregel (ψ, φ) ,

$$s(\psi, \varphi) = \int_{\mathcal{X}} r_x(\psi, \varphi) f(x) dx, \text{ so ist auch } s \text{ in } (\psi_0, \delta_0)$$

minimal. Offenbar haben $r_x(\psi, \varphi)$ und $s(\psi, \varphi)$ eine ähnliche Form wie die entsprechenden Risiken in sequentiellen Entscheidungsverfahren. Wird dort (vgl. [3], [45], [12]) z.B. die Anzahl der Merkmale schrittweise vergrößert, so wird hier die Klassenanzahl bzw. nach Ordnen der $\Gamma \in \mathcal{X}_d$ nach $J(\Gamma)$ der Informationsgehalt schrittweise verringert. Dabei gibt uns die MBE die Möglichkeit, für jede Beobachtung $x \in \mathcal{X}$ zunächst (durch ψ_0) die Klassen zu bestimmen, die günstig getrennt

werden können, und anschließend (durch δ_0) die optimale Zuordnung durchzuführen. Jedem $x \in \mathcal{X}$ wird also ein Paar (Γ, J) mit $J \in \Gamma$ zugeordnet. Setzt man

$\tilde{\Omega} = \{(\Gamma, J) \in \mathcal{A} \times \mathcal{P}(I) : J \in \Gamma\}$ ($\mathcal{P}(I)$ sei die Potenzmenge von I ohne die leere Menge) und $L^\Gamma(j, J) = L^\Gamma(J', J)$ für $j \in J' \in \Gamma$, $J \in \Gamma$, erhält man

1.2 SATZ

Sei $\tilde{L}: I \times \tilde{\Omega} \rightarrow \mathbb{R}$ definiert durch $\tilde{L}(j, (\Gamma, J)) = L^\Gamma(j, J) + \gamma c(\Gamma)$ für $j \in I, (\Gamma, J) \in \tilde{\Omega}$. Dann ist $\tilde{\delta}: \tilde{\Omega} \times \mathcal{X} \rightarrow [0, 1]$ mit

$$\tilde{\delta}((\Gamma, J) | x) = 1 \iff \psi_0(\Gamma | x) = 1 \text{ und } \delta_0^\Gamma(J | x) = 1$$

eine Bayes-Entscheidung zu $(I, \tilde{\Omega}, \tilde{L})$ und es gilt

$$\tau_x(\tilde{\delta}) = r_x(\psi_0, \delta_0) \quad \forall x \in \mathcal{X} \quad \text{und} \quad s(\tilde{\delta}) = s(\psi_0, \delta_0).$$

Beweis: Für jede Entscheidungsfunktion $\tilde{\varphi}: \tilde{\Omega} \times \mathcal{X} \rightarrow [0, 1]$ ist

$$\begin{aligned} \tau_x(\tilde{\varphi}) &= \sum_{j \in I} \sum_{(\Gamma, J) \in \tilde{\Omega}} \tilde{L}(j, (\Gamma, J)) \gamma_x(j) \tilde{\varphi}((\Gamma, J) | x) \\ &= \sum_{j \in I} \sum_{(\Gamma, J) \in \tilde{\Omega}} (L^\Gamma(j, J) + \gamma c(\Gamma)) \gamma_x(j) \tilde{\varphi}((\Gamma, J) | x). \end{aligned}$$

Sei jetzt $\tilde{\delta}((\Gamma, J) | x) = 1$ für $(\Gamma, J) \in \tilde{\Omega}$, also

$$\begin{aligned} \tau_x(\tilde{\delta}) &= \sum_{j \in I} (L^\Gamma(j, J) + \gamma c(\Gamma)) \gamma_x(j) = \sum_{j \in I} L^\Gamma(j, J) \gamma_x(j) + \gamma c(\Gamma) \\ &= r_x(\psi_0, \delta_0) \leq \sum_{j \in I} (L^{\Gamma'}(j, J') + \gamma c(\Gamma')) \gamma_x(j) \end{aligned}$$

für alle $(\Gamma', J') \in \tilde{\Omega}$ nach Konstruktion von ψ_0 und δ_0 . Somit

wird $\tau_x(\varphi)$ für jede nicht randomisierte Entscheidungsfunktion

φ mit $\varphi((\Gamma', J') | x) = 1$ und damit auch für jede randomisierte

Entscheidungsfunktion (da $\tilde{\Omega}$ endlich) durch $\tau_x(\tilde{\delta})$ minimiert

und es gilt $\tau_x(\tilde{\delta}) = r_x(\psi_0, \delta_0)$ und $s(\tilde{\delta}) = s(\psi_0, \delta_0)$. ■

Vergleicht man das Risiko $s(\psi_0, \delta_0)$ mit dem der nach dem Differenzenkriterium (I, 4.3) besten Klasseneinteilung $\bar{\Gamma}$, ergibt sich $s(\bar{\Gamma}) + \gamma c(\bar{\Gamma}) = s(\bar{\psi}, \delta_0) > s(\psi_0, \delta_0)$. Dabei ist $\bar{\psi}(\Gamma|x) = 1$ für alle $\Gamma = \bar{\Gamma}$ und alle $x \in \mathcal{X}$ und stellt die beste unter allen auf $\mathcal{X}$ konstanten Stoppfunktionen dar. Wegen der Größe von $\mathcal{K}_{\mathcal{X}}$ ist das Auffinden von $\bar{\Gamma}$ und $(\bar{\psi}_0, \delta_0)$ mit erheblichem Aufwand verbunden. Um die Anzahl der bei der Klassifizierung in Frage kommenden Klasseneinteilungen zu reduzieren, betrachten wir ab jetzt nur noch solche Verlustfunktionen, die (L1) bis (L3) aus Kap. I, 5 genügen. Dies gilt u.a. für L_1 mit $L_1(J) = 1 \quad \forall J \in \mathcal{R}(I)$ als auch für L_2 mit $L_2(J) = \max_{j \in J} L_j$. Aus (L1)-(L3) folgt nun eine besonders einfache Form von τ : Ist $\delta^\Gamma(J|x) = 1$, so gilt $L^\Gamma(J) \gamma_x^\Gamma(J) = \max_{J' \in \Gamma} L^\Gamma(J') \gamma_x^\Gamma(J')$ und $\tau(\Gamma|x) = \sum_{\Gamma' \ni J+J} L^\Gamma(J') \gamma_x^\Gamma(J')$. Damit kann man $\tilde{\mathcal{U}}$ auf $\mathcal{R}(I)$ beschränken:

1.3 SATZ

Sei $\mathcal{K} = \{\Gamma \in \mathcal{K}_0 : \Gamma = \Gamma_J = \{J, \{k\}; k \neq J, J \in \mathcal{R}(I)\}\}$.

Ist $p_j L_j \neq 0 \quad \forall j \in I$, gilt für alle $x \in \mathcal{X}$

(1) $\psi_0(\Gamma|x) = 0$ für $\Gamma \in \mathcal{K}_0 \setminus \mathcal{K}$

(2) Ist für $\Gamma = \Gamma_J \in \mathcal{K}$ J mindestens zweielementig ($|J| \geq 2$) und ist $\psi_0(J|x) := \psi_0(\Gamma_J|x) = 1$, so folgt $\delta^\Gamma(J|x) = 1$ und

$$\tau(J|x) := \tau(\Gamma|x) = \sum_{k \neq J} L_k \gamma_x(k).$$

Beweis: (1) Sei $\Gamma = \{I_1, \dots, I_M\}$ mit $|I_1|, |I_2| \geq 2$ und o.B.d.A.

$\delta^\Gamma(I_1|x) = 1$ mit $1 \neq 2$. Für $\Gamma' = \{I_1, I_2, \dots, I_M, \{j\}; j \in I_2\}$ folgt aus (L1) und (L2) $\sum_{j \in I_2} L_j \gamma_x(j) \leq L(I_2) \sum_{j \in I_2} \gamma_x(j)$ und

$$\begin{aligned} \tau(\Gamma|x) &= \sum_{k=1}^M L(I_k) \gamma_x^\Gamma(I_k) > \sum_{k=2,1}^M L(I_k) \gamma_x^\Gamma(I_k) + \sum_{j \in I_2} L_j \gamma_x(j) \\ &= \tau(\Gamma'|x). \end{aligned}$$

Andererseits ist $J(\Gamma) < J(\Gamma')$, also

$$r(\Gamma|x) = r(\Gamma|x) + \gamma c(\Gamma) > r(\Gamma'|x) + \gamma c(\Gamma') = r(\Gamma'|x)$$

und damit $\psi_0(\Gamma|x) = 0$.

(2) Sei jetzt $\Gamma = \Gamma_J$ und $|J| \geq 2$.

$$\text{Ist } \delta^\Gamma(k_0|x) = 1 \text{ und } k_0 \notin J, \text{ gilt } \tau(\Gamma_0|x) = \sum_{j \neq k_0} L_j \gamma_x(j) \leq$$

$$\sum_{j \neq k_0} L_k \gamma_x(k) + L(J) \sum_{j \in J} \gamma_x(j) = \tau(\Gamma|x).$$

Wieder folgt aus $c(\Gamma_0) = 0 < c(\Gamma)$ $r(\Gamma_0|x) < r(\Gamma_J|x)$. Dies widerspricht aber $\psi_0(\Gamma_J|x) = 1$. $\blacksquare$

Setzt man $\tilde{\mathcal{O}}_0 = \{(\Gamma, J) \in \tilde{\mathcal{O}} : \Gamma \in \mathcal{K} \text{ mit } \Gamma = \Gamma_0, \text{ falls } |J|=1\}$,

so ist $J \mapsto \begin{cases} (\Gamma_J, J) & , \text{ falls } |J| \geq 2 \\ (\Gamma_0, \{i\}) & , \text{ falls } J = \{i\} \end{cases}$ eine bijektive

Abbildung von $\mathcal{P}(I)$ auf $\tilde{\mathcal{O}}_0$. Dabei ist $\tilde{\delta}((\Gamma, J)|x) = 0$ für $(\Gamma, J) \notin \tilde{\mathcal{O}}_0$. Für $(\Gamma_J, J) \in \tilde{\mathcal{O}}_0$ und $\tilde{\delta}((\Gamma_J, J)|x) = 1$ schreibe $\tilde{\delta}(x) = J$.

Eine Ausdehnung der Menge der Aktionen von I auf $\mathcal{P}(I)$ hat auch CHU[7] durchgeführt. Dabei hat er jedoch nur Verlustfunktionen $\tilde{L}$ behandelt, für die

$$\tilde{L}(j, J) = \begin{cases} a & , \text{ falls } j \notin J \\ a_k & , \text{ falls } j \in J \text{ und } |J|=k \end{cases} \text{ gilt.}$$

Es wird also der Verlust einer Fehlentscheidung immer gleich mit a gewichtet, während der Verlust durch eine fehlende Unterscheidung zwischen den Klassen aus J nur von der Anzahl der Klassen aus J abhängig gemacht wird. Dieses Verfahren ist den medizinischen Entscheidungen aber weniger angemessen,

da doch selbst bei gleicher Klassenanzahl der Verzicht auf die Unterscheidung großer oder wichtiger Klassen schwerer wiegt als der kleiner oder unbedeutender Klassen.

Da beim Übergang von Γ_J zu $\Gamma_{J \cup \{k\}}$ das a posteriori Risiko um $L_k \gamma_x(k)$ sinkt und der Informationsverlust um $c(J \cup \{k\}) - c(J)$ (es sei $c(J) := c(\Gamma_J)$ und $r(J|x) := r(\Gamma_J|x)$) wächst, kann man $\tilde{\alpha}_0$ weiter einschränken:

1.4 FOLGERUNG

Sei $p(J_1) \leq p(J_2)$ und sei $k \notin J_1 \cup J_2$. Dann folgt aus $r(J_1|x) \leq r(J_2|x)$ auch $r(J_1 \cup \{k\}|x) \leq r(J_2 \cup \{k\}|x)$ für alle $x \in \mathcal{X}$.

Beweis: Neben $r(J_1|x) - r(J_1 \cup \{k\}|x) = L_k \gamma_x(k) = r(J_2|x) - r(J_2 \cup \{k\}|x)$ gilt nach I, 3.4 $r(J_1) - r(J_1 \cup \{k\}) \leq r(J_2) - r(J_2 \cup \{k\})$, weil $p(J_1) \leq p(J_2)$ ist, also auch:

$$\begin{aligned} & r(J_1 \cup \{k\}|x) - r(J_2 \cup \{k\}|x) \\ &= r(J_1|x) - r(J_2|x) + \gamma(c(J_1) - c(J_2)) + \gamma((c(J_1 \cup \{k\}) - c(J_1)) - (c(J_2 \cup \{k\}) - c(J_2))) \\ &\leq r(J_1|x) - r(J_2|x) \leq 0. \quad \blacksquare \end{aligned}$$

Behandelt man also die Zerlegungen Γ_J mit kleinem $p(J)$ zuerst, kann man frühzeitig terminale Zuordnungen in bestimmte größere Klassen ausschließen. Ist etwa bei $N=4$ $p_1 \leq p_2 \leq p_3 \leq p_4$ und $r(1|x) \leq r(2|x)$, müssen $\{2,3\}$, $\{2,4\}$ und $\{2,3,4\}$ nicht mehr untersucht werden.

Über die Anzahl der Elemente kann man $\mathcal{P}(I)$ vollständig ordnen: $J_1 < J_2$, falls $|J_1| < |J_2|$ und für $J_1 = \{i_1 < \dots < i_M\}$ und $J_2 = \{j_1 < \dots < j_M\}$, falls $i_1 < j_1$ oder $i_1 = j_1, \dots, j_k = i_k$ und $i_{k+1} < j_{k+1}$ für ein $1 \leq k \leq M-1$. Um Folgerung 1.4 anzuwenden, ist die Ordnung $\mathcal{L}$ geeignet ($J_1 \mathcal{L} J_2 \Leftrightarrow p(J_1) < p(J_2)$ oder $p(J_1) = p(J_2)$ und $J_1 < J_2$).

Den Bezug zur Reject-Regel stellt her

1.5 LEMMA

(1) Für alle $\Gamma \in \mathcal{K}_0$ gilt: $\psi_0(\Gamma|x)=1 \Rightarrow \tau(\Gamma|x) \leq \gamma \quad \forall \quad x \in \mathcal{X}$

(2) Für $\gamma > (N-1)/N$ ist $\psi_0(I|x) = 0$.

Beweis: (1) $r(\Gamma|x) = \tau(\Gamma|x) + \gamma c(\Gamma) \leq r(I|x) = 0 + \gamma = \gamma$.

(2) Da für $\Gamma = \{I_1, \dots, I_M\}$ $\tau(\Gamma|x) \leq (M-1)/M \leq (N-1)/N$ gilt,

ist $r(I|x) = \gamma > (N-1)/N \geq \tau(\Gamma_0|x) = r(\Gamma_0|x)$.

2. Modifizierte Reject-Entscheidung

Wie für das Reject-Verfahren kann man jetzt auch für die modifizierte Bayes-Entscheidung Reject- und Akzeptiermengen einführen. Für $\psi \in \Psi(\mathcal{X})$ sei

$$\mathcal{V}_R(\psi) = \{x: \psi(I|x) \neq 0\} \quad \text{und} \quad \mathcal{V}_A(\psi) = \{x: \psi(I|x) = 0\}$$

Dabei ist nach 1.5 $\mathcal{V}_R(\psi_0)$ für $(N-1)/N < \gamma$ leer. Da für $x \in \mathcal{V}_A(t)$ bei $\gamma = t$ $\tau(x) = \tau(\Gamma_0|x) < t = \tau(I|x)$, also $\psi_0(I|x)=0$ gilt, hat man stets $\mathcal{V}_A(\psi_0) \supseteq \mathcal{V}_A(t)$ und $\mathcal{V}_R(\psi_0) \subseteq \mathcal{V}_R(t)$.

Die MBR zu $\gamma = t$ ergibt also in jedem Fall geringere Reject-Raten. Dabei werden mögliche totale Zurückweisungen in modifizierte Entscheidungen umgewandelt und es wird ein großes Maß an Information bei gleichzeitiger Beschränkung des Fehlerkennungsrisikos erhalten (s.u: Beispiel 2.6).

Offenbar ist jedoch die Suche der optimalen Entscheidung mit großem Rechenaufwand verbunden, müssen doch stets alle $r(J|x)$ berechnet und verglichen werden (für $J \in \mathcal{P}(I)$). Um dieses Verfahren abzukürzen, wird nun die Stoppregele ψ_t eingeführt.

Dazu sei $\mathcal{P}(I)$ nach der Größe des Informationsgehalts vollständig geordnet: $J_1 \overset{J}{<} J_2$, falls $J(J_1) < J(J_2)$ oder $J(J_1) = J(J_2)$ und $J_1 > J_2$ (mit $J(J) = J(\Gamma_J)$).

Betrachtet man die a posteriori Risiken mit sinkendem Informationsgehalt, so stoppt Ψ_t bei dem ersten $J \in \mathcal{I}(I)$, für das $\tau(J|x) < t(J) := t \cdot J(J)/J$ gilt (für $t > 0$), also

$$\Psi_t(J|x) = \Psi_t(\Gamma_J|x) = 1 \Leftrightarrow \begin{cases} \tau(J'|x) \geq t(J') \text{ für alle } J' \supset J \text{ und} \\ \tau(J|x) < t(J) \end{cases}$$

Damit gilt sogar $V_A(t) = \{x: \Psi_t(\Gamma_0|x) = 1\}$. Somit wird $V_R(t)$ durch Ψ_t in (echt) modifizierte Entscheidungen und einen totalen Reject zerlegt. (Ψ_t, δ_0) heißt modifizierte Reject-Entscheidung (MRE) (für $\delta_0 \in \mathcal{D}(\mathcal{X})$). Betrachtet man nur die Ψ_0 , die im Zweifel zurückweisen ($\Psi_0(I|x) = 1 \Leftrightarrow r(r|x) \geq t \forall r \in \mathcal{X}$), bekommt man

2.1 FOLGERUNG

Sei Ψ_0 die Stoppfunktion zu $\gamma = t$ und Ψ_t wie oben definiert.

Dann ist $\mathcal{R}_R(\Psi_t) = \mathcal{R}_R(\Psi_0)$.

Beweis: $x \in \mathcal{R}_R(\Psi_t) \Leftrightarrow \tau(J|x) \geq t(J) \quad \forall J \in \mathcal{I} \Leftrightarrow$

$$\tau(J|x) + t \cdot c(J) \geq t = \tau(I|x) \Leftrightarrow x \in \mathcal{R}_R(\Psi_0) \quad \square$$

Für den Informationsgehalt von $\Psi \in \Psi(\mathcal{X})$,

$$J(\Psi) = \int_{\mathcal{X}} \left(\sum_{\Gamma \in \mathcal{X}} \Psi(\Gamma|x) J(\Gamma) \right) f(x) dx,$$

erhält man dann $J(\Psi_0) \leq J(\Psi_t)$ für alle $t > 0$. Definiert man

$$\mathcal{R}(\Psi) = \int_{\mathcal{X}} \Psi(I|x) f(x) dx, \text{ die modifizierte Rejectwahrscheinlichkeit,}$$

$$\mathcal{E}(\Psi) = \int_{\mathcal{X}} \left(\sum_{\Gamma \in \mathcal{X}} \Psi(\Gamma|x) \tau(\Gamma|x) \right) f(x) dx, \text{ das modifizierte}$$

Fehlerkennungsrisiko von Ψ , folgt $\mathcal{R}(\Psi_t) = \mathcal{R}(\Psi_0)$ und

$$\mathcal{E}(\Psi_t) \geq \mathcal{E}(\Psi_0), \text{ außerdem } t \geq s(\tilde{s}) = t \mathcal{R}(\Psi_0) + \mathcal{E}(\Psi_0) + \frac{t(J - J(\Psi_0))}{J}$$

und im Vergleich mit der gewöhnlichen Bayes-Entscheidung

2.2 Satz

Sei $\mathcal{S}$ das Bayes-Risiko zu $(I, \zeta, \tilde{J})$ mit $\Omega = \mathcal{O} = I$. Dann gilt

$$0 \leq \frac{J - J(t)}{J} \leq \frac{s - \mathcal{E}(t)}{t} \quad \text{mit } J(t) = J(\Psi_t) \text{ und } \mathcal{E}(t) = \mathcal{E}(\Psi_t).$$

Beweis:

$$g - E(t) \geq \int \sum_{x \in \mathcal{X}} \sum_{\Gamma \in \mathcal{U}} \psi_t(\Gamma|x) (\tau(\Gamma_0|x) - \tau(\Gamma|x)) f(x) dx$$

$$\geq t \int \sum_{x \in \mathcal{X}} \sum_{\Gamma \in \mathcal{U}} \psi_t(\Gamma|x) c(\Gamma) f(x) dx,$$

da für $\psi_t(\Gamma|x)=1$ ($\Gamma \neq \Gamma_0$) $\tau(\Gamma_0|x) \geq t$ und $\tau(\Gamma_0|x) \leq J(\Gamma)t/J$,
also $\tau(\Gamma_0|x) - \tau(\Gamma|x) \geq t(1 - J(\Gamma)/J) = t c(\Gamma)$ gilt.

$$\text{Andererseits hat man } \int \sum_{x \in \mathcal{X}} \sum_{\Gamma \in \mathcal{U}} \psi_t(\Gamma|x) c(\Gamma) f(x) dx$$

$$= \int \sum_{x \in \mathcal{X}} \sum_{\Gamma \in \mathcal{U}} \psi_t(\Gamma|x) (1 - J(\Gamma)/J) f(x) dx = (J - J(t))/J \quad \text{und}$$

damit die Behauptung. $\square$

Da bei ψ_t das Verfahren gestoppt wird, wenn $\tau(J|x)$ erstmalig eine vorgegebene Schranke unterschreitet, entspricht es dem "sequential probability ratio test", wie er von FU [12] in Verallgemeinerung des Verfahrens von WALD [44] konstruiert worden ist. Dieser Test wird u.a. dazu benutzt, die Merkmale nacheinander, abhängig von variablen Stoppgrößen, in die Klassifikation einzuführen. Wie die sequentielle Merkmalsauswahl mit den modifizierten Entscheidungsverfahren effektiv kombiniert werden kann, soll in §3 erläutert werden. Zunächst werden noch einige Eigenschaften von ψ_0 und ψ_t gezeigt:

2.3 LEMMA

Für $j \in I$ und $J \in \mathcal{P}(I)$ gilt

(1) Aus $j \notin J$, $p(\tilde{\delta}(x)) \geq p(J \cup \{j\})$ und $L_j \tilde{\gamma}_x(j) < t(c(J \cup \{j\}) - c(J))$
folgt $j \notin \tilde{\delta}(x)$

(2) Aus $J(\tilde{\delta}(x)) \leq J(J)$ und $L_j \tilde{\gamma}_x(j) \geq t(J)$ folgt $j \in \tilde{\delta}(x)$
und $j \in J_t$ mit $\psi_t(J_t|x)=1$.

Beweis:

(1) ang. $j \in \tilde{\delta}(x) = J'$. Dann folgt $L_j \gamma_x(j) = \tau(J' \setminus \{j\} | x) - \tau(J' | x) > \tau(c(J') - c(J' \setminus \{j\})) > \tau(c(J \setminus \{j\}) - c(J))$, weil nach Satz 3.4 aus Kap I $c(J \setminus \{j\}) - c(J)$ monoton wachsend mit $p(J)$ ist.

(2) ang. $j \notin \tilde{\delta}(x) = J'$. Dann gilt $\tau(J' | x) > L_j \gamma_x(j) > \tau(J(J)/J > \tau(J(J')/J$. ■

Damit hat man die Möglichkeit, schrittweise abzuklären, ob j der terminalen Entscheidungsmenge angehört. Sind etwa die Ungleichungen in 2.3(1) schon für kleine $p(J)$ erfüllt, kann man auf eine Untersuchung aller J mit größerem $p(J \setminus \{j\})$ verzichten. Ähnlich verhält es sich mit 2.3(2). Hier kann man bei großem $L_j \gamma_x(j)$ schon frühzeitig erkennen, ob j der terminalen Entscheidungsmenge angehört.

2.4 SATZ

Sei $j \in I$ und $q_j = \min_{k \neq j} p_k$. Dann ist

$L_j \gamma_x(j) > t$ hinreichend und $L_j \gamma_x(j) > \frac{t}{J} (p_j + q_j) J (\frac{p_j}{p_j + q_j}, \frac{q_j}{p_j + q_j})$ notwendig für $j \in \tilde{\delta}(x)$.

Beweis: Mit $|J|=1$, also $\Gamma_J = \Gamma_0$ und $t(J)=t$ folgt die hinreichende Bedingung aus 2.3(2), die notwendige folgt aus 2.3(1), weil

$$\frac{p_j + q_j}{J} - J(\frac{p_j}{p_j + q_j}, \frac{q_j}{p_j + q_j}) = \min_{J \in \mathcal{P}(I), j \notin J} (c(J \setminus \{j\}) - c(J)) \text{ gilt. } \blacksquare$$

Da für $\tilde{\delta}(x)=I$ $L_j \gamma_x(j) > \frac{t}{J} J(p_j, 1-p_j)$ für alle $j \in I$ gilt, ist die Bedingung $\forall x \in X \quad \exists j \in I: L_j \gamma_x(j) < \frac{t}{J} J(p_j, 1-p_j)$

hinreichend dafür, daß $\mathcal{V}_R^0(\varphi_0) = \mathcal{V}_R^0(\varphi_t) = \emptyset$ ist. Für großes N reduziert sich diese Ungleichung nach I, 3.3 auf $L_j f_j(x) < t f(x)$.

2.5 BEMERKUNGEN

(1) Man kann auch in ähnlicher Weise eine modifizierte Reject-Regel konstruieren, die auf jeder Stufe eine Zurückweisungs - möglichkeit vorsieht. Dabei kann man für jedes $x \in \mathcal{X}$ wieder den Reject-Verlust in Abhängigkeit der tatsächlichen Klassen - zugehörigkeit durch K_j wichten, indem man $r_K(\Gamma|x) = \sum_{j \in I} K_j \gamma_x(j) + K(x)c(\Gamma)$ mit $K(x) = \sum_{j \in I} K_j \gamma_x(j)$, $\gamma_K(\Gamma|x) = \min(\gamma(\Gamma|x), K(x))$ über alle $\Gamma \in \mathcal{K}$ minimiert (vgl. II,1).

Da für $\Gamma \neq \Gamma_0$ und $\gamma_K(\Gamma|x) = K(x)$ $r_K(\Gamma|x) > r_K(\Gamma_0|x) = K(x) = r_K(\Gamma_I|x)$ ist, ist ein Reject auf der Stufe $\Gamma \neq \Gamma_0$ immer schlechter als ein Reject bei $\Gamma = \Gamma_0$ und dieser ist gleich gut wie eine Zu - ordnung $\tilde{\delta}(x) = I$.

Also entspricht dieses Verfahren einer MBE mit Wichtung $K(x)$ (statt γ). Ähnlich wie in 1.2 kann man hierfür zeigen, daß diese Regel einer Bayes-Entscheidung $\tilde{\delta}_K: \tilde{\mathcal{O}}_K \times \mathcal{X} \rightarrow [0,1]$ entspricht. Dabei ist $\tilde{\mathcal{O}}_K = \tilde{\mathcal{O}} \cup (\mathcal{K} \times \{0\})$ und $\tilde{L}_K(j, (\Gamma, J)) = L^\Gamma(j, J) + K_j c(\Gamma)$ für $(\Gamma, J) \in \tilde{\mathcal{O}}$, $L_K(j, (\Gamma, 0)) = K_j(1 + c(\Gamma))$ für $(\Gamma, 0) \in \mathcal{K} \times \{0\}$. Ebenso kann eine MRE (ψ_K, δ_0) mit $\psi_K: \mathcal{K} \times \mathcal{X} \rightarrow [0,1]$ konstruiert werden.

(2) Eine ähnliche Entscheidungsregel wie MBE und MRE, mit $\tilde{\mathcal{O}} = \mathcal{P}(I) \cup \{0\}$, haben QUESENBERRY und GESSAMAN[37] entwickelt. Sie konstruierten Entscheidungsregionen $S(J)$ ($I \neq J \in \mathcal{P}(I)$) und $S(0)$ mit Hilfe von Toleranzgebieten $A_1, \dots, A_N \subset \mathcal{X}$:

$$S(J) = \bigcap_{j \in J} \bar{A}_j \cap \bigcap_{k \notin J} A_k \text{ und } S(0) = \bigcap_{j \in I} A_j \cup \bigcap_{j \in I} \bar{A}_j.$$

Dabei entspricht $S(0)$ der Rejectmenge. In [37] war das Ziel, Entscheidungsregionen $S(J)$ bzw. Mengen A_j zu finden, die die

bedingten Wahrscheinlichkeiten einer falschen Zuordnung

$$q_j = P_j(Q_j) = \int_{Q_j} f_j(x) dx \quad \text{mit } Q_j = \bigcup_{J \in \mathcal{R}(I), j \neq J+I} S(J)$$

beschränkten und die bedingten Rejectwahrscheinlichkeiten

$P_j(S(o))$ minimierten. Für $N=2$ gelang dies mit Hilfe des Satzes von Neyman-Pearson. Dabei war für geeignete $c'_1, c'_2 \in \mathbb{R}$ $A_1 = \{x: f_1(x) \leq c'_1 f_2(x)\}$ und $A_2 = \{x: f_1(x) > c'_2 f_2(x)\}$.

Setzt man für beliebiges N $A_j = \{x: f_j(x) < c_j f(x)\}$, bekommt man auch $q_j \leq c_j$ und für die Stoppfunktion ψ mit

$$\psi(\Gamma_J | x) = 1 \Leftrightarrow x \in S(J) \quad \tau(J|x) \leq \sum_{k \neq J} p_k L_k c_k.$$

Für $c_k = t$, $L_k = 1$ und $p_k = 1/N$ $\forall k \in I$ wird daraus

$$\tau(J|x) \leq t(N-m)/N \sim t \mathbb{I}(\Gamma_J)/\mathbb{I} = t(J) \text{ für } |J| = m \text{ und } N \text{ gegen } \infty$$

nach dem Satz aus I, 3.2. In diesem Fall genügt $\tau(J|x)$ also derselben Schranke, gleichgültig ob man nach ψ , ψ_o oder ψ_t stoppt. Ähnliche Bedingungen findet man auch für die Reject-Mengen:

Da für $x \in \bigcap_{j \in I} A_j$ $f(x) < f(x) \sum_{j \in I} c_j p_j = \bar{c} f(x)$ und

für $x \in \bigcap_{j \in I} \bar{A}_j$ $1 \geq \bar{c}$ gilt, ist

$$S(o) = \begin{cases} \bigcap_{j \in I} A_j & \Leftrightarrow 1 < \bar{c} \\ \bigcap_{j \in I} \bar{A}_j & \Leftrightarrow 1 \geq \bar{c} \end{cases}.$$

Meistens wird man $\bar{c} \leq 1$ finden. Dann ist also $S(o) = \emptyset$,

wenn für alle $x \in \mathcal{X}$ $f_j(x) < c_j f(x)$ für ein $j \in I$ gilt. Dies entspricht für $L_j = 1$ und $c_j = t$ nun wieder der im Anschluß an 2.4 angegebenen Bedingung für $\gamma_R(\psi_t) = \gamma_R(\psi_o) = \emptyset$.

Da nach Vorgabe von $c_1, \dots, c_N$ die $S(J)$ sehr leicht zu bestimmen sind, ist dieses verallgemeinerte Verfahren von QUESENBERY und GESSAMAN immer ohne großen Rechenaufwand durchführbar.

2.6 BEISPIEL

Seien $f_j \sim \mathcal{N}(M_j, \Sigma_j)$ für $j=1, \dots, N$. Wie in II, 2.3 zerlegen wir $\mathcal{X} = \bigcup_{i,j} S_{ij}$, so daß $S_{ij} \in \{x: p_i f_i(x), p_j f_j(x) > p_k f_k(x) \quad \forall K \neq i, j\}$.

Wir nehmen nun an, daß für $K \neq i, j$ die $p_k f_k(x)$ annähernd konstant (gleich $a_{ij}^{(k)}$) auf S_{ij} sind. Dies ist im univariaten Fall besonders bei großen Varianzen σ_k erfüllt.

Sind m_1 bzw. m_2 die jeweils auf $\mathcal{X}$ größten bzw. zweitgrößten $p_j f_j(x)$, so ist $x \in V_R(t) \cap S_{ij} \Leftrightarrow f(x)/m_1(x) \leq 1/(1-t)$.

$$\frac{f(x)}{m_1(x)} = 1 + \frac{m_2(x)}{m_1(x)} + \frac{a_{ij}}{m_1(x)} \quad (a_{ij} = \sum_{k \neq i, j} a_{ij}^{(k)})$$

ist maximal für alle $x_{ij} \in G_{ij} = \{x \in S_{ij}: p_i f_i(x) = p_j f_j(x)\}$.

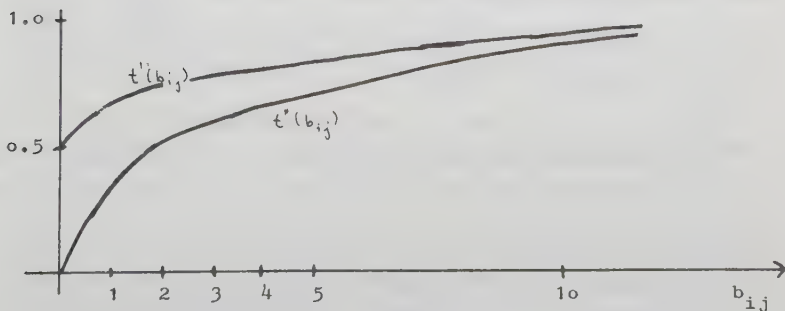
Also ist $V_R(t) \cap S_{ij} = \emptyset$ genau dann, wenn

$$\frac{f(x_{ij})}{m_1(x_{ij})} = 2 + b_{ij} \leq \frac{1}{1-t} \quad (\text{mit } b_{ij} = a_{ij}/m_1(x_{ij}))$$

oder wenn $t \geq t' := \frac{b_{ij}+1}{b_{ij}+2}$. Analog bekommt man

$$V_R(t | \Gamma_{\{i,j\}}) = \emptyset \Leftrightarrow t \geq t'' := \frac{b_{ij}}{b_{ij}+2}.$$

Dabei ist immer $0 \leq b_{ij} \leq N-2$. Graphisch erhält man



Man sieht, daß für kleine b_{ij} die Hinzunahme einer modifizierten Entscheidung $\tilde{s}(x) = j$ schon für $|j| = 2$ große

Änderungen an der Rejectwahrscheinlichkeit bewirkt.

Bemerkt sei, daß die Forderung $f_k \sim \text{const.}$ auf S_{ij} ($k \neq i, j$) durch Beschränkungen des Wachstums der f_k ersetzt werden kann.

Für univariate Normalverteilungen $f_j \sim \mathcal{N}(\mu_j, \sigma)$ mit gleicher Varianz kann man die $b_{ij}^{(k)} = f_k(x_{ij})/m_1(x_{ij})$ leicht berechnen:

$$b_{ij}^{(k)} = \frac{f_k}{m_1}(\frac{\mu_k + \mu_j}{2}) = \exp(-(\mu_k - \mu_j)(\mu_k + \mu_j)/(2\sigma^2)) .$$

Ist $\mu_1 \leq \dots \leq \mu_N$, muß man nur den Fall $j=i+1$ betrachten (da sonst $S_{ij} = \emptyset$). Dann ergibt sich für $b_{ij}^{(k)}$ ($k > j$):

$(\mu_k - \mu_j)/\sigma$	$(\mu_j - \mu_1)/\sigma$	1	2	3	5
1		e^{-1}	$e^{-1.5}$	e^{-2}	e^{-3}
2		e^{-3}	e^{-4}	e^{-5}	e^{-7}
3		$e^{-6} \sim 10^{-2.6}$	10^{-3}	10^{-4}	10^{-5}
5		10^{-6}	10^{-7}	10^{-9}	10^{-11}
10		10^{-23}	10^{-25}	10^{-28}	10^{-32}

Über die Äquivalenz $p_i f_i(\bar{x}) = (1-t)f(\bar{x}) \Leftrightarrow \frac{p_i f_i(\bar{x})}{p_1 f_1(\bar{x})} = \frac{t}{1-t} + b_{ij}$

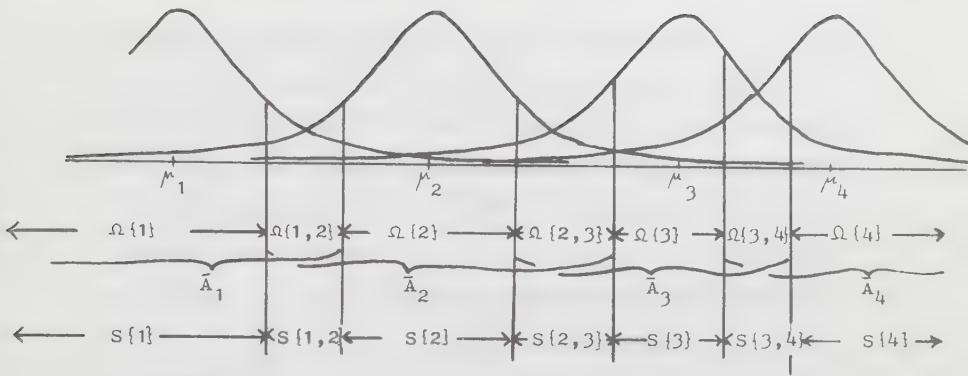
kann man nun auch, wie in II, 3.4, die Entscheidungsgrenzen auf S_{ij} berechnen.

Sind die $b_{ij} \sim b$ ($\forall i < j$), wird für $t > t^*(b)$ die modifizierte Reject-Entscheidung äquivalent der von QUESENBERY und GESSAMAN (bei $p_1 = \dots = p_N$):

Mit $c_j = t - (1-t)b$ ist nämlich $m_2(x) = c_j f(x) \Leftrightarrow m_1(x) = (1-t)f(x)$.

Damit stimmen die resultierenden Entscheidungsregionen

$\Omega(J)$ und $S(J)$ überein:



2.7 BEMERKUNG

Die modifizierten Entscheidungsregeln beinhalten auch das neueste Verfahren in der automatischen medizinischen Diagnostik, das der "fuzzy diagnosis" (SMETS et al. [40]). Hier werden zunächst die einzelnen Krankheiten nach dem Schweregrad in Subklassen unterteilt. Anschließend wird angegeben, mit welcher Wahrscheinlichkeit der beobachtete Wert einer dieser Krankheiten zuzuordnen ist und ob er mit hinreichender Genauigkeit (etwa 95%) in eine der Subklassen einzuordnen ist. Dieses Verfahren entspricht ungefähr dem folgenden:

- (1) Prüfe, wie fein sich jede Klasse in Subklassen unterteilen läßt (Diskussion der BKs und von J).
- (2) Führe für die so definierte Klasseneinteilung eine MRE durch (etwa mit $t=0.05$).

Wie SMETS et al. bemerkten, ist das Verfahren der "fuzzy diagnosis" flexibler und dem üblichen ärztlichen Diagnoseverhalten angemessener als das übliche Bayes-Verfahren.

Durch die hier angegebene Handhabung kann man nun die Sub -
klassen sinnvoll definieren und bei der Klassifikation unter
Beschränkung des Fehlerkennungsrisikos einen großen Teil an
Information erhalten. Darüberhinaus ergeben sich aus den Er -
örterungen von HABBEMA [17], daß eine Entscheidung vom
modifizierten Typ auch Hinweise auf das gleichzeitige Vor -
liegen mehrerer Krankheiten geben kann.

3. Modifizierte Entscheidung mit sequentieller Merkmalauswahl

In allen Mustererkennungsprozessen ist die Auswahl geeigneter
Merkmale von großer Bedeutung. Dieses Problem, das von
vielen Autoren aus den verschiedensten Blickwinkeln be -
trachtet worden ist ([12],[23],[28],[31],[32],[33],[45])
und das schon in I,7.3 aufgegriffen wurde, war jedoch nicht
primärer Inhalt dieser Arbeit. Ich möchte aber noch einige
Hinweise geben, wie die hier erzielten Ergebnisse möglicher -
weise nutzbringend auf die Merkmalsauswahl angewendet werden
können. Dazu will ich ein Verfahren vorstellen, das folgendes
gewährleistet:

- (1) Das a posteriori Risiko jeder Entscheidung bleibt stets
durch eine vorgegebene Zahl beschränkt.
- (2) Eine variable Zuordnung im Sinne der "fuzzy diagnosis"
wird ermöglicht.
- (3) Die Auswahl der Merkmale wird successive, in Abhängig -
keit der jeweils zuvor beobachteten Werte vorgenommen.

Dazu sei wieder $I = \{1, \dots, N\}$ die Menge der möglichen Zustände
des (analogen) Musters X - als solche sind EKg-Muster Ele -
mente des unendlich-dimensionalen Vektorraums $C(\mathbb{R})$ - und

$J = \{1, \dots, s\}$ die Menge der in Frage kommenden zu messenden Merkmale. Außerdem seien $t_1, \dots, t_s > 0$ vorgegeben.

Für alle $I' \subseteq I$ und $J' \subseteq J$ seien die Näherungen von $g^{J'}(I')$ (Bayes-Risiko zur Klassenauswahl I' und zur Merkmalsauswahl J') bestimmt. Dazu ist (nach I, 2.5) die Berechnung von $\phi^{J'}(i, j)$ für alle $J' \subseteq J$ und $i, j \in I$ notwendig.

1. Schritt

a) Auswahl des 1. Merkmals

$$g^{\{1\}}(I) = \min_{j \in J} \{g^{\{j\}}(I)\} \quad (\text{nach Umnummerierung})$$

b) Entscheide $\tilde{\delta}_{t_1}(x_1) = I_1$ (MRE zu I bzgl. t_1).

Bilde $J_1 = \{1\}$.

⋮

k-ter Schritt

J_{k-1} sei die Menge der $k-1$ beobachteten Merkmale. Außerdem sei $x^{(k-1)} = (x_1, \dots, x_{k-1})$ der Vektor der beobachteten Werte (nach Umnummerierung) und $\tilde{\delta}_{t_{k-1}}(x^{(k-1)}) = I_{k-1}$.

a) Auswahl des k-ten Merkmals

$$\text{Minimiere über alle } j \in J \setminus J_{k-1} \quad g^{J_{k-1} \cup \{j\}}(I_{k-1}) + z(j) \\ (\text{bzw. a'}) \quad \sum_{i \in I_{k-1}} \sum_{x^{(k-1)}} (1) \sum_{i \neq 1 \in I_{k-1}} g^{J_{k-1} \cup \{j\}}(i, 1) + z(j))$$

Das minimierende j ist das k-te Merkmal.

b) Entscheide $\tilde{\delta}_{t_k}(x^{(k)}) = I_k$ (MRE zu I_{k-1} bzgl. t_k).

Bilde $J_k = J_{k-1} \cup \{k\}$ (nach Umnummerierung)

Stoppe, falls $I_k = I_{k-1}$.

Das Verfahren wird also gestoppt, wenn das jeweils beste Merkmal die Unterscheidungskraft nicht weiter verbessert. (hier kann man auch sicherheitshalber noch solche Merkmale mit ähnlich guten Kriterien a) untersuchen). Am Ende steht eine

modifizierte Entscheidung $\tilde{S}(x^{(k)}) = I_k$.

In die Größe $z(j)$, die Kosten der Untersuchung des j -ten Merkmals, gehen dabei in medizinischen Diagnoseverfahren neben den materiellen Kosten auch die Risiken des Patienten ein (mögliche Nebenwirkungen, z.B. durch Röntgenbestrahlung oder invasive Methoden).

Ein Problem dürfte die Berechnung der $\binom{N}{2}(2^S-1)$ Bhattacharyya-Koeffizienten sein. Für die Ekg-Klassifizierung von PIPBERGER mit 7 Klassen und 66 Merkmalen sind das 10^{21} Größen, für die in [31] ausgewählten 16 Merkmale 10^6 Größen. Sind die Merkmale unabhängig voneinander bzgl. aller Dichten $f_1, \dots, f_N$ (was selten der Fall ist), vereinfacht sich die Situation erheblich: Dann folgt aus $f_i(x_1, \dots, x_r) = \prod_{u=1}^r f_i^{(u)}(x_u)$, $1 \leq r \leq s$ für $J' = \{1, \dots, r\}$

$\phi^{J'}(f_i, f_j) = \prod_{u=1}^r \phi^{(u)}(f_i^{(u)}, f_j^{(u)})$, so daß nur die Speicherung von $\binom{N}{2} \cdot s$ Größen notwendig ist.

Auch der "Generalized Sequential Probability Ratio Test" (GSPRT) von REED (siehe FU[12]) verringert auf jeder Stufe der Merkmalsauswahl die Menge der für die Klassifizierung in Frage kommenden Klassen. Dabei wird aber stets eine Einordnung in $k \leq I$ verlangt. Die Tatsache, daß die Reihenfolge der Merkmalsauswahl für die Dauer des Prozesses von besonderer Bedeutung ist (vgl. [12]), wird in dem oben angegebenen Verfahren doppelt berücksichtigt, zum einen durch die jeweilige Konzentration auf I_k , zu anderen durch die Wichtung mit $\gamma_{x^{(k)}}(i)$ in a'). Dabei dürfte die Abhängigkeit der Merkmalsauswahl von den zuvor beobachteten Werten von Nutzen sein. Dies be -

sonders dann, wenn der Aufwand für die Messung und Berechnung der $x_1, \dots, x_s$ sehr groß ist. Dagegen wird in allen anderen mir bekannten Verfahren zur Merkmalsauswahl diese anhand durchschnittlicher Kriterien von vorneherein festgelegt. Für die Ekg-Diagnostik ist das von MUCCIARDI und GOSE [33] diskutiert worden.

Liste der Bezeichnungen

ϕ : S.8f.
 γ : S.63; Γ , Γ_0 , $N(\Gamma)$, $\Omega(\Gamma)$ etc.: S.2, Γ_j : S.66
 δ : S.2; δ_K : S.48; δ_0 : S.64; $\tilde{\delta}$: S.65; $\Delta(J)$: S.5
 $\varepsilon(J)$: S.3; $\varepsilon(\Gamma)$: S.6
 Υ , $\Upsilon_x(j)$: S.1; Υ' : S.27f.
 $g(\varphi|j)$, $g(\varphi)$: S.1; g : S.2; $g(J)$: S.3; g_j : S.5; $g(\psi, \varphi)$: S.64
 $\tau_x(\varphi)$: S.1; $\tau_x(i)$: S.48; $\tau(x)$: S.49; $\tau(\Gamma|x)$: S.63; $\tau(J|x)$: S.65
 ψ : S.1
 Ψ , $\Psi(\lambda_x)$: S.63; $\Psi_0(J|x)$: S.66; Ψ_t : S.70
 Ω : S.1; Ω_i : S.2; $\Omega_i(\Upsilon)$: S.31

 $\mathcal{O}$: S.1; $\tilde{\mathcal{O}}_0$: S.67
BK-L: S.9
 $\mathcal{D}$, $\mathcal{D}^*$, $\mathcal{D}_0^*$: S.1; d_u, d_1 : S.9; "D-besser": S.24; $\mathcal{D}(\mathcal{K}_u)$: S.63
 $e_x(\varphi)$: S.47; E : S.48f.
 $\mathcal{F}$, f_j, f : S.1
H: S.17
J: S.17f.
 $K, K(x)$: S.47; $\mathcal{K}_u$: S.63
L: S.1; $L(J), \bar{L}$: S.27, $L(x)$: S.50
 $m(x)$: S.50
 $\mathcal{N}(\mathcal{C}_j, \mu_j)$: S.7; $\mathcal{N}(M_j, \Sigma_j)$: S.12
 p_j : S.1; $p(J)$: S.5; p_j^{Γ} : S.27; $P(t)$: S.53
"Q-besser": S.24; $\mathcal{Q}_1$: S.32
 $r_x(\varphi)$: S.47; $R(\varphi)$: S.48; $R(L)$: S.49; $r_x(\psi, \varphi)$: S.63
 $v_R(L), v_A(L)$: S.49
 $x, \mathfrak{X}$: S.1

Daneben gilt immer $S = S(\Gamma_0)$, $J = J(\Gamma_0)$, $N = N(\Gamma_0)$ etc.

Neben den mathematischen Symbolen $\forall$ = "für alle" und

$\exists$ = "es existiert" sind noch erklärt:

$|\Sigma| = \det(\Sigma)$: S.12; x' = Transponierte von x : S.12;

$\bar{A} = \mathcal{K} \setminus A$: S.4; $P_j(A) = P(A|j)$: S.31; $\Gamma \prec \Gamma_0$: S.2;

$\mathbb{K}$: S.68; $\mathbb{Z}$: S.69

LITERATURVERZEICHNIS

- [1] ANDERSON, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley, New York 1958
- [2] ASH, R.: Information Theory. Wiley, New York 1967
- [3] BLACKWELL, D., GIRSHICK, A.: Theory of Games and Statistical Decisions. Wiley, New York 1954
- [4] CHOW, C.K.: An optimum character recognition system using decision functions. IRE Trans. Electronic Computers, Dec. 1957, 247-254
- [5] CHOW, C.K.: On optimum recognition error and reject tradeoff. IEEE Trans. Inform. Theory, IT-16, 41-46 (1970)
- [6] CHU, J.T.: Optimal decision functions for computer character recognition. J. Ass. for Comp. Machinery, Vol.12, No.2, 213-226 (1965)
- [7] CHU, J.T.: Some decision making methods applicable to the medical sciences. Proc. 5-th Berkeley Symposium on Math. Stat. and Prob., 1967. Vol.IV, pp. 925-934
- [8] CHU, J.T., CHUEH, J.C.: Error probability in decision functions for character recognition. J. Ass. for Comp. Machinery, Vol.14, No.2, 273-280 (1967)
- [9] COVER, T.M.: Rates of convergence for nearest neighbor procedures. Proc. 1st Ann. Hawaii Conf. on Systems Theory (Jan. 1968)
- [10] COVER, T.M., HART, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inform. Theory, IT-13, 21-27 (1967)
- [11] DAS GUPTA, S.: Theories and methods in classification: A review. In CACOULOS, T.: Discrimination Analysis and Applications. Academic Press, New York 1973, pp. 77-137
- [12] FU, K.S.: Sequential Methods in Pattern Recognition and Machine Learning. Academic Press, New York 1968

- [13] FUKUNAGA, K., KRILE, T.F.: Calculation of Bayes recognition error for two multivariate Gaussian distributions. IEEE Trans. Computers C-18, 220-229 (1969)
- [14] GOOD, I.J., CARD, W.I.: The diagnostic process with special reference to errors. Meth. Inform. Medicine 10, 176-188 (1971)
- [15] GREGG, W.D., HANCOCK, J.C.: An optimum decision directed scheme for Gaussian mixtures. IEEE Trans. Inform. Theory IT-14, 451-461 (1968)
- [16] GREGOR, J.: The decomposition of a distribution into Gaussian components (Algol program). Computer Programs in Biomedicine, Vol. 7, 251-255 (1977)
- [17] HABBEEMA, J.D.F.: Models for diagnosis and detection of combinations of diseases. In DOMBAL, F.T., GREMY, F. (Eds): Decision Making and Medical Care. Can Information Science Help?, pp. 399-410. North Holland Publ. Comp., Amsterdam 1976
- [18] HELLMAN, M.E.: The nearest neighbor classification rule with a reject option. IEEE International Symp. on Inform. Theory, Holland. Program and Abstracts of Papers, Noordwijk 1970
- [19] HERMANS, J., EGGERMONT, J.J., HAGEDOOREN, J., ODENTHAL, D.W.: Probabilistic differential diagnosis of auditory dysfunction on the basis of electrocochleography. Meth. Inform. Medicine 14, 87-95 (1975)
- [20] HESS, C.F., BRODDA, K., WELLNER, U.: On the construction of optimum categories in biomedical data recognition problems. Erscheint in Konferenzbericht zu Medical Informatics Berlin, Sept. 1979. Springer Verlag, Heidelberg
- [21] HEYDORN, R.P.: An upper bound estimate on classification error. IEEE Trans. Inform. Theory 783-784 (Sept. 1968)
- [22] HUZURBASAR, V.S.: Exact forms of some invariants for distributions admitting sufficient statistics. Biometrika, Vol. 42, 533-537 (1955)

- [23] KAILATH, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. Communi - cation Technology, Vol. Com-15, No.1, 52-60 (1967)
- [24] KOBAYASHI, H., THOMAS, J.B.: Distance measures and re - lated criteria. Proc. Ann. Conf. Circuit and System Theory, 5th, Allerton, Oct. 1967, pp 491-500
- [25] KULLBACK, S.: Information Theory and Statistics. Wiley, New York 1959
- [26] LACHENBRUCH, P.A.: Some unsolved problems in discriminant analysis. In DOMBAL, F.T., GREMY, F. (Eds): Decision Making and Medical Care. Can Information Science Help?, pp 423-432. North Holland Publ. Comp., Amster - dam 1976
- [27] LAINIOTIS, D.G.: A class of upper bounds on probability of error for multihypotheses pattern recognition. IEEE Trans. Inform. Theory: 730-731 (Nov. 1969)
- [28] MARILL, T., GREEN, D.M.: On the effectiveness of receptors in recognition systems. IEEE Trans. Inform. Theory: 11-17 (Jan. 1963)
- [29] MATUSITA, K.: Classification based on distance in multivariate Gaussian cases. Proc. 5th Symp. Math. Stat. and Prob., Univ. Calif. 1967, Vol. I: 299-304
- [30] MC QUEEN, J.: Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symp. Math. Stat. and Prob., Vol. I: 281-297 (1967)
- [31] MICHAELIS, J.: Zur Anwendung der Diskriminanzanalyse für die medizinische Diagnostik. Habilitationsschrift Universität Mainz 1972
- [32] MIN, P.J., LANDGREBE, D.A., FU, K.S.: On feature se - lection in multiclass pattern recognition. Proc. 2nd Princeton Conf. Inform. Sciences and Systems, March 1968, pp 453-457
- [33] MUCCIARDI, A.N., GOSE, E.E.: A comparison of seven techniques for choosing subsets of pattern recognition properties. IEEE Trans. Computers, Vol. C-20: 1023-1030 (1971)

- [34] NORUSIS, M.J., JACQUES, J.A.: Diagnosis II. Diagnostic models based on attribute clusters: A proposal and comparisons. *Comp. Biomed. Res.* 8: 173-188 (1975)
- [35] PETERSON, D.W.: Some convergence properties of a nearest neighbor rule. *IEEE Trans. Inform. Theory*, IT-16: 26-31 (1970)
- [36] PIPBERGER, H.V., KLINGEMAN, J.D., COSMA, J.: Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain. *Meth. Inform. Med.* 7, 79-92 (1968)
- [37] QUESENBERRY, C.P., GESSAMAN, M.P.: Nonparametric discrimination using tolerance regions. *Ann. Math. Stat.*, Vol. 39, No. 2, pp 664-673 (1968)
- [38] SCHWEPPE, F.: On the Bhattacharyya distance and the divergence between Gaussian processes. *Information and Control* 11: 373-395 (1967)
- [39] SEBESTYEN, G.: *Decision - making Processes in Pattern Recognition*. Mac Millan, New York 1962
- [40] SMETS, P., VAINSEL, H., BERNARD, R., KORNREICH, F.: Bayesian Probability of fuzzy diagnosis. In SHIRES, D.B., WOLF, H. (Eds): *MEDINFO 77*, pp 121-122. North Holland Publ. Comp., Amsterdam 1977
- [41] TERRENOIRE, M., TOUNISSOUX, D.: Pseudoquestionnaires and information. In DOMBAL, F.T., GREMY, F. (Eds): *Decision Making and Medical Care. Can Information Science Help?*, pp 413-420. North Holland Publ. Co., Amsterdam 1976
- [42] VIRNICH, H., WINKLER, C.: Das Bayes'sche Theorem in der medizinischen Diagnostik. *Meth. Inform. Med.* 15: 211-216 (1976)
- [43] WAGNER, G., TAUTU, P., WOLBER, U.: Problems of medical diagnosis - a bibliography -. *Meth. Inform. Med.* 17, No. 1, 55-74 (1978)

- [44] WALD, A.: Statistical Decision Functions. Wiley, New York 1950
- [45] YOUNG, T.Y., CALVERT, T.W.: Classification, Estimation and Pattern Recognition. Am. Elsevier Publ. Co., New York, 1974.
- [46] YOUNG, T.Y., CORALUPPI, G.: Stochastic estimation of a mixture of normal density functions using an information criterion. IEEE Trans. Inform. Theory IT-16: 258-263 (1970)

LEBENS LAUF

Am 9.3.1951 wurde ich, Friedrich Heß, als drittes von fünf Kindern des Arztes Dr. Anton Heß und seiner Ehefrau, der Ärztin Margarethe Heß, geb. Loenartz, in Koblenz geboren. Nach dem Besuch der Volksschule (1957 - 1961) und des neu - sprachlichen Gymnasiums (1961 - 1969) in Cochem bestand ich dort 1969 das Abitur.

Im WS 1969/70 begann ich an der Universität Mainz das Studium der Mathematik mit Nebenfach Physik. Im WS 1971/72 legte ich die Diplom-Vorprüfung und im WS 1975/76 die Diplom-Hauptprüfung im Fach Mathematik ab. Seit dem SS 1975 studiere ich an der Universität Mainz das Fach Humanmedizin, in dem ich im WS 1976/77 die ärztliche Vorprüfung bestand.

Mainz, den 28.12.1979

Friedrich Heß



Druck: D. Gräbner, 8602 Altendorf - Tel. 09545 / 8791